

# DSC 423 Group Project

## Milestone 5: Group Project Proposal

### Ronaldlee Ejalu, Pinki Sharma, Kajal Patel, Bansari Gandhi

Health care, one of the largest industries is much affected by the ongoing pandemic as they strive to control it. Due to this pandemic many people are losing their lives and family members. Day by day, the situation is getting worse; moreover, people are losing their hope to cope up with the situation. The number of COVID-19 deaths are increasing in a rapid phase and especially, when we look through the USA COVID-19 dataset, we found there are several factors contributing to this such as the different age groups, gender, the underlying disease conditions and different geographical states of the USA.

#### Section 2: Data Preparation

Two data sets were acquired from the CDC and Kaggle website. The conditions contributing to deaths involving coronavirus disease 2019 was collected from <https://healthdata.gov/dataset/conditions>. This dataset consists of the conditions contributing to the Covid-19 Death cases in the United States of America based on the state. The data about the states was collected from <https://www.kaggle.com/omer2040/usa-states-to-region?select=states.csv>. This data set has all the different states of the USA corresponding to the different geographical regions.

Deaths are those occurring among confirmed cases on the day of the death. The two data sets, COVID and states, were combined to create a single data set. Among the 12 variables, the number of COVID-19 deaths is the dependent variable. Region, Condition Group, condition, sex and age are categorical variables. Number of COVID-19 deaths is numerical. Each row is a data object corresponding to a condition which has an impact on the number of COVID-19 deaths broken down by age groups and the different states where these deaths are occurring. Region, Condition Group, condition, sex and age were converted into factors with different levels. Number of COVID-19 deaths and Sex has some entries, which marked as "NA" and this needs to be cleaned up.

The code snippet below shows a dim Function in R with our data frame as an object argument to it returning the number of columns and rows.

```
> dim(ConditionscontributingCOVID19US)
[1] 12290    10
```

Our data set had 12290 conditions contributing to deaths involving coronavirus disease 2019

with 11 variables. There are some noisy variables, which don't have any impact on the dependent variable. These are flag, ICD10\_codes, Start.Week, End.Week and Date.as.of. These variables were removed using the following process:

```
> ConditionscontributingCOVID19US$Flag <- NULL
> dim(ConditionscontributingCOVID19US)
[1] 12290    9
> ConditionscontributingCOVID19US$data.as.of <- NULL
> dim(ConditionscontributingCOVID19US)
[1] 12290    8
>> dim(ConditionscontributingCOVID19US)
[1] 12290    7
```

To get a better idea of the distribution of variables in the dataset, we used a summary function as below:

```
> summary(ConditionscontributingCOVID19US)
  Data.as.of           Start.Week          End.Week
Length:12290      Length:12290      Length:12290
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


State           Condition.Group       Condition
Length:12290      Length:12290      Length:12290
Class :character   Class :character   Class:character
Mode  :character   Mode  :character   Mode  :character


ICD10_codes     Age.Group
Length:12290      Length:12290
Class :character   Class :character
Mode  :character   Mode  :character


Number.of.COVID.19.Deaths
Min. : 0.0
1st Qu.: 0.0
Median : 20.0
Mean   : 320.8
3rd Qu.: 98.0
Max.   :191211.0
NA's   : 3116
```

In R NA (Not Available) is used to represent missing values, so our data set had 3116 missing

values in the Number.of.COVID.19.Deaths variable. These either had to be replaced or removed. To get a better understanding of the missing values in the Number.of.COVID.19.Deaths variable, we created a subset of data only by extracting out those rows where this variable had missing values using both the filter() function from the dplyr package and is.na logical comparison operator as shown below: These observations may not have any impact on the dependent variable of our model, however, when you disregard cases with any missing variables, you lose useful information that the no missing values convey.

```
> RowsWhereNumberOfDeathIsNA <- ConditionscontributingCOVID19US  
%>% filter(is.na(Number.of.COVID.19.Deaths))
```

The Number.of.COVID.19.Deaths variable missing values (NA's) were replaced with the average mean of this variable values, excluding the missing values and those values were assigned to a newly created variable, NoOfCovid19Deaths as part of the dataset.

```
> library(plyr)  
> library(dplyr)  
> ConditionscontributingCOVID19US$NoOfCovid19Deaths <-  
replace(ConditionscontributingCOVID19US$Number.of.COVID.19.Deat  
hs,  
is.na(ConditionscontributingCOVID19US$Number.of.COVID.19.Deaths  
),  
mean(ConditionscontributingCOVID19US$Number.of.COVID.19.Deaths,  
na.rm = TRUE))
```

Also, the dplyr package in r is one of the most popular packages for data manipulation. The Plyr package is split apply combine paradigm for R. Before using the replace function, which modifies the existing data, this package had to be attached and use a library() function to attach those packages in my current Rstudio session.

Now that we have two variable names with the Number of COVID 19Deaths.That is Number.of.COVID.19.Deaths and NoOfCovid19Deaths variables. We dropped the Number.of.COVID.19.Deaths variable since the variable name was constructed with character ‘.’ instead of underscores. This variable name has a bad naming convention, so we set it to NULL so that it is out of our data set.

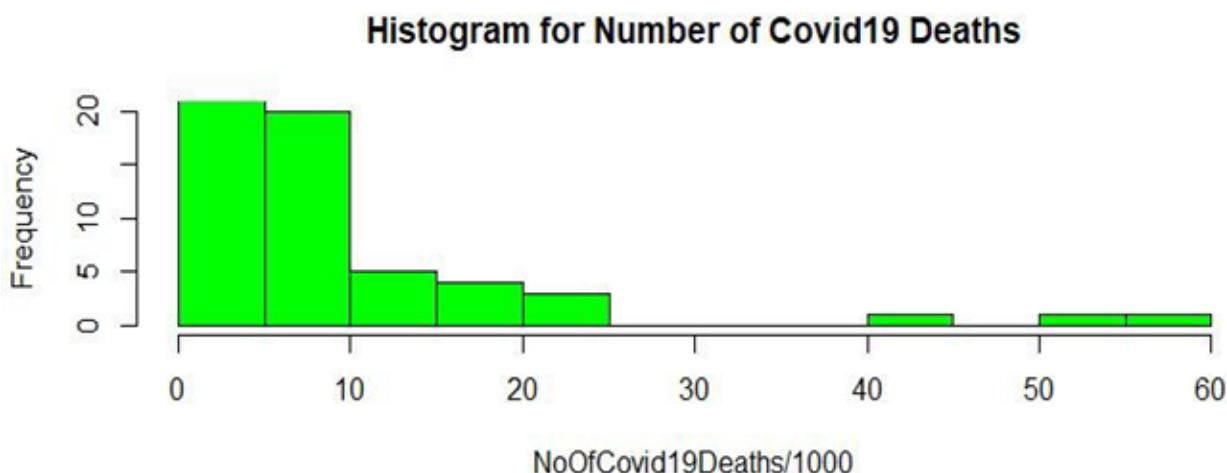
```
>ConditionscontributingCOVID19US$Number.of.COVID.19.Deaths <-  
NULL
```

We summarized and aggregated our data set by condition group, condition and age group. Running the aggregate() function to get the data into subsets, we assigned the result of the summary dataset into SummarizeddataSet.

```
> SummarizeddataSet <-  
aggregate(x=ConditionscontributingCOVID19US$NoOfCovid19Deaths,  
by=list("ConditionGroup"=ConditionscontributingCOVID19US$Condition.Group,  
"Condition" =  
ConditionscontributingCOVID19US$Condition, "AgeGroup" =  
ConditionscontributingCOVID19US$Age.Group), FUN=sum)
```

The aggregate function splits the data into subsets, computes summary statistics for each and returns the result in a convenient form.

### Histograms showing the distribution of relevant numeric variables -



The above histogram shows a visual representation of the number of Covid19 deaths. As defined above there is a big percentage of Covid19 deaths that have occurred within a range of 0 to ten thousand since the pandemic began. Because of the large numerical values in the number of COVID19 deaths in our dataset, on the x axis, we have the number of COVID19 deaths divided by 1000. Yes, this number seems to be out of the roof given the fact that there has been loss of lives. Looking at the histogram, it is right skewed, we can predict that data is normally not distributed. There are outliers on the right part of the histogram. This indicates that there were measurement errors in our data set.

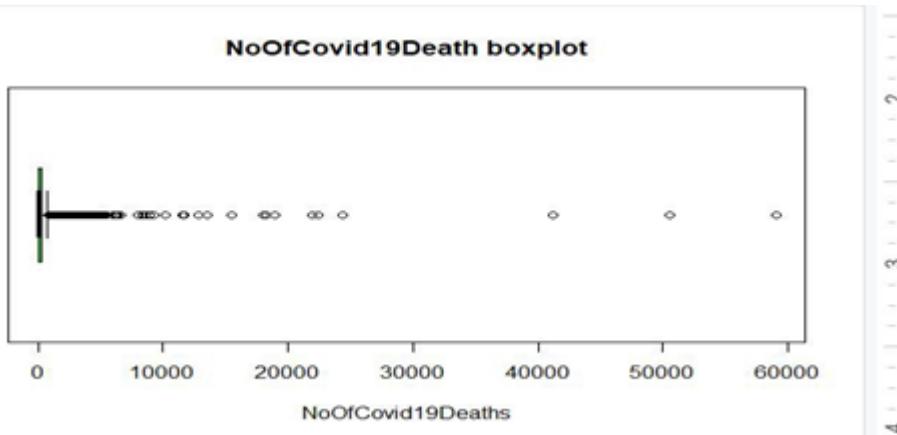
## **Five number Summaries showing the center and the spread -**

```
> summary(Summarizeddataset$NoofCovid19Deaths)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	0.0	58.0	320.8	320.8	191211.0

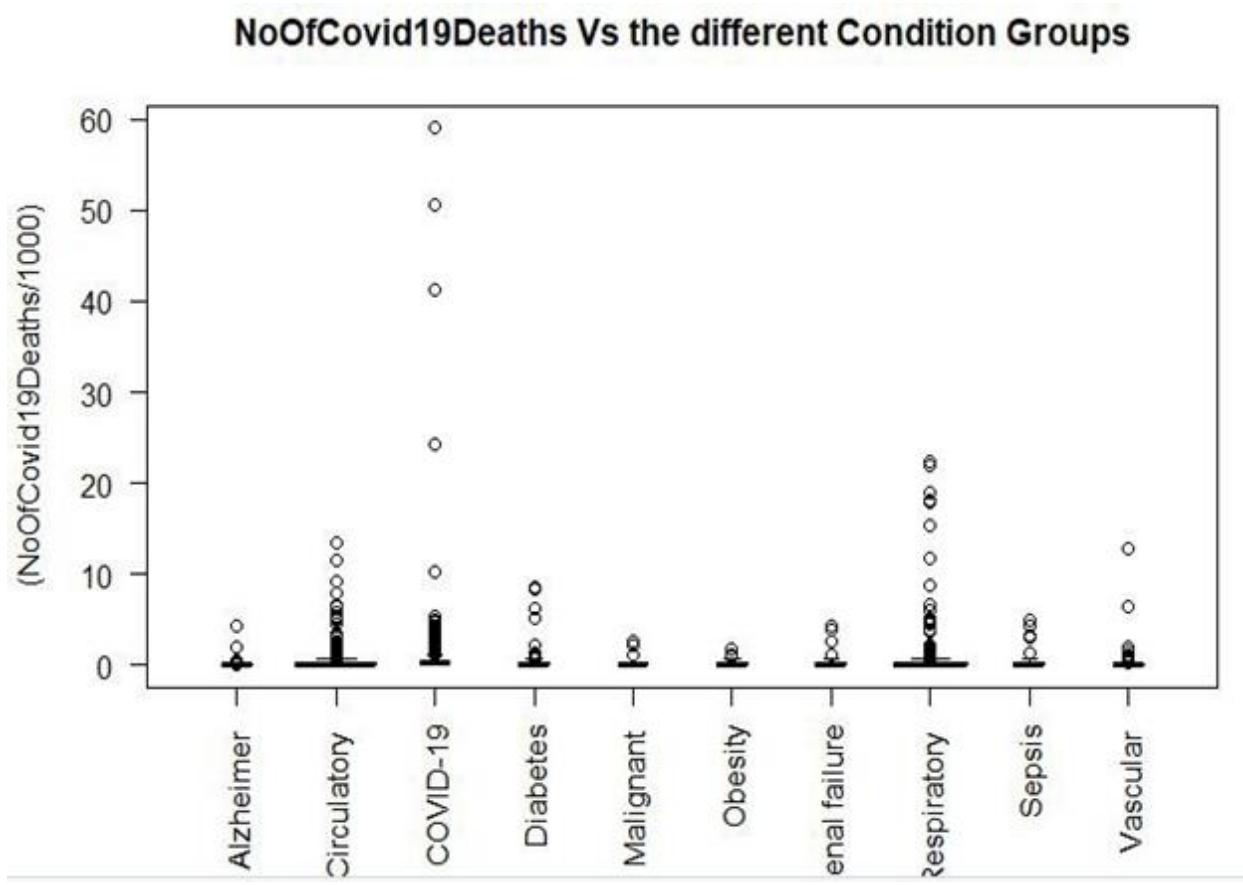
The summary function demonstrated the distribution of the number of COVID19 deaths variable. The five-number summary for the number of COVID19 deaths had a minimum value of 0 and maximum value of 191211. The mean value was 320.8, which showed that the data was skewed. It has been found that the mean was greater than the median, one could interpret that there is a positive skew of the distribution that means number of COVID19 deaths was unevenly distributed.

## **Box plots showing the distributions of variables over two or more classes -**



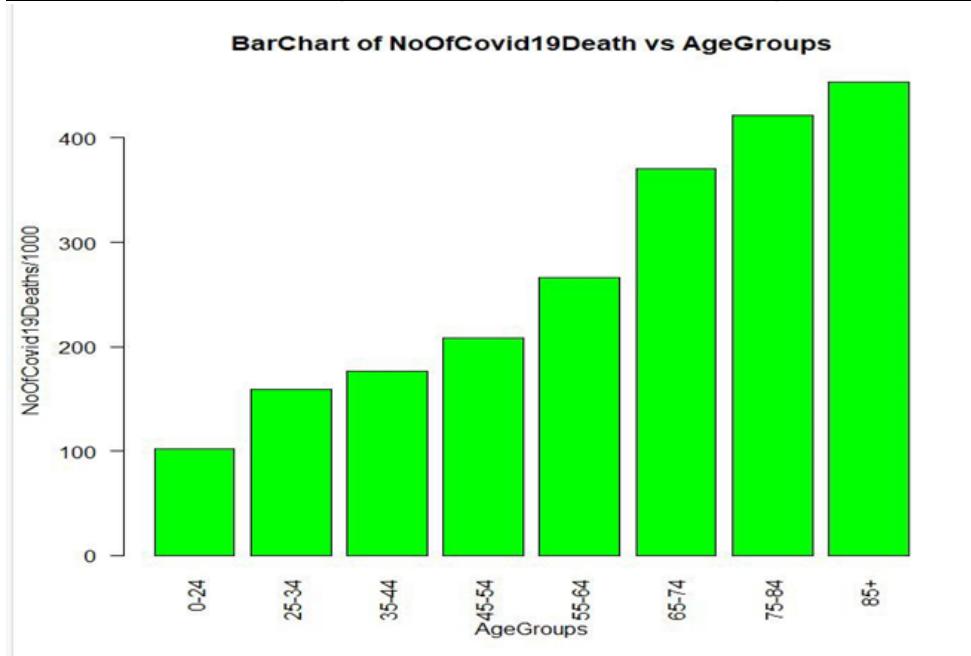
Box plot is the graphic display of five number summary, after we looked at it, it was clear that number of COVID19 deaths (NoOfCovid19Deaths) was unevenly distributed and it was positively skewed towards the right. We see that they were outliers ( data points that are located outside the whisker of the box plot) on the right most part of the boxplot.

## Comparing the distribution of NoOfCovid19Deaths to the different Condition groups -



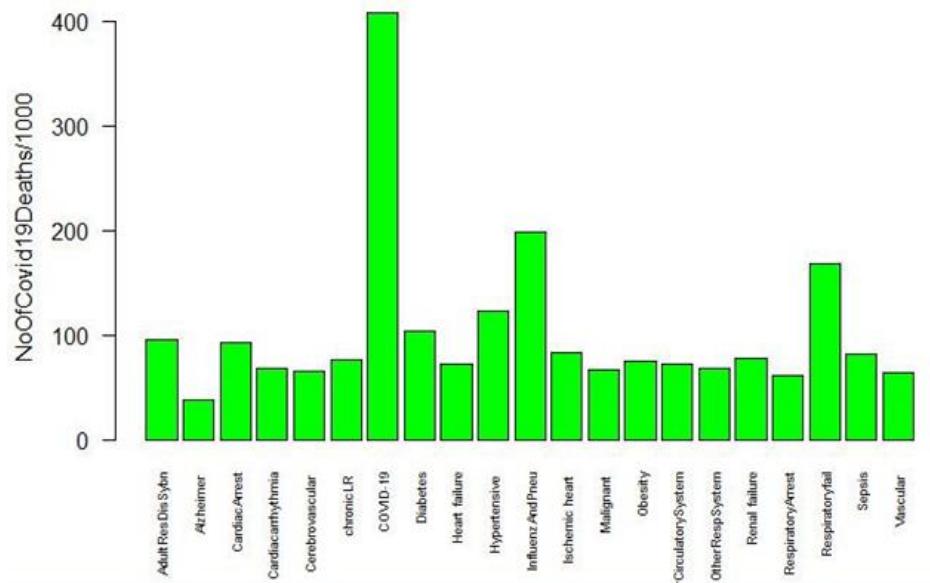
The Boxplot above shows how different groups of conditions are responsible for the number of Covid-19 Deaths. Looking through the box plot we observed that COVID-19 has the highest number of deaths outside its range, that is to say the number of outliers. Because of the large numerical values in the number of COVID19 deaths in our dataset, on the y axis, we have the number of COVID19 deaths divided by 1000. In addition to this, Respiratory, circular, vascular, diabetes diseases also had outliers.

## **Bar charts showing the distribution of categorical variables -**



Here, we used a vertical bar graph to describe NoOfCovid19Deaths using different age groups. We used AgeGroups as a categorical variable on the X-axis and used NoOfCovid19Deaths/1000 as a numerical variable on the Y-axis. Because of the large numerical values in the number of COVID19 deaths in our dataset, on the y axis, we have the number of COVID19 deaths divided by 1000. NoOfCovid19Deaths/1000 for age group 0-24 are lowest amongst all the age groups. After that, it increases as age-groups increases. Also, it shows a sudden increase in NoOfCovid19Deaths for age group 65-74 and continues to rise for remaining age groups too. For age group of 85+, deaths are very high amongst all the age-groups. Overall, as age-group increase, NoOfCovid19Deaths increased as well. That means the death ratio was highest among old people and lowest among young people.

**BarChart of NoOfCovid19Death vs Conditions**



Here, the bar graph represented NoOfCovid19Deaths versus Conditions. We used Conditions as a categorical variable on the X-axis to describe the various diseases that patients might have contracted before being tested for COVID19 and used NoOfCovid19Deaths/1000 as a numerical variable on the Y-axis. Because of the large numerical values in the number of COVID19 deaths in our dataset, on the y axis, we have the number of COVID19 deaths divided by 1000. As shown above, COVID-19 accounts for most deaths compared with other conditions given their complexities. Also, Pneumonia and Respiratory failure are one of the conditions responsible for the high number of Covid19 deaths. Overall, the above graph shows fluctuation in the number of covid19 deaths divided by 1000 (NoOfCovid19Deaths/1000) with respect to different conditions.

Our observation was number of covid19 deaths per 1000 was highest for COVID-19 and lowest for Alzheimer disease conditions.

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9
dNoOfCovidperAge	0-24	25-34	35-44	45-54	55-64	65-74	75-84	85
102179.611837802	102179.6110	0	0	0	0	0	0	0
159580.097122302	0	159580.09710	0	0	0	0	0	0
176046.924787443	0	0	176046.9240	0	0	0	0	0
208171.896664487	0	0	0	208171.8960	0	0	0	0
266721.032701112	0	0	0	0	266721.032700	0	0	0
370530.818181818	0	0	0	0	0	370530.818180	0	0
421847.914323087	0	0	0	0	0	0	421847.914	0
453256.603662525	0	0	0	0	0	0	0	453256.6

The table above defined a Pivot table of the various Age-groups that have died from Covid19.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	NoOfCO	Adult res	Alzheimer	Cardiac ar	Cardiac ar	Cerebrov	Chronic lo	COVID-19	Diabetes	Heart fail	Hyperten	Influenza	Ischemic	Malignant	Obesity	Other disa	Other disa	Renal fail
1	37538.41	0	37538.41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	61176.61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	63655.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	65731.3	0	0	0	0	65731.3	0	0	0	0	0	0	0	0	0	0	0	0
5	65998.36	0	0	0	0	0	0	0	0	0	0	0	0	65998.36	0	0	0	0
6	68318.41	0	0	0	68318.41	0	0	0	0	0	0	0	0	0	0	0	0	0
7	68528.67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	68528.67	0	0
8	72556.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	72556.25	0	0
9	72705.06	0	0	0	0	0	0	0	0	72705.06	0	0	0	0	0	0	0	0
10	74549.45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	74549.45	0	0
11	76255.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	78112.41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78112.41
13	81093.24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	82553.16	0	0	0	0	0	0	0	0	0	0	0	0	82553.16	0	0	0	0
15	93259	0	0	93259	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	95684.47	95684.47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	103627	0	0	0	0	0	0	0	103627	0	0	0	0	0	0	0	0	0
18	122590.4	0	0	0	0	0	0	0	0	0	122590.4	0	0	0	0	0	0	0
19	168345.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	198437.4	0	0	0	0	0	0	0	0	0	0	0	0	198437.4	0	0	0	0
21	407618.7	0	0	0	0	0	0	0	407618.7	0	0	0	0	0	0	0	0	0

The table above defined a Pivot table showing the number of Conditions contributing to the NoOfCovid19Deaths.

We divided the COVID data set into different datasets based on the different age groups of 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; each of the group members were assigned to work on a particular group to design and implement a multiple regression model to predict the number of COVID19 deaths. The model implemented would either be a first order model or interaction model. The final model will have to be validated by N fold cross validation.

**Sub section author: Pinki Sharma**

**Sub section author student id: 1983149**

**Based on M5, M7 and M8**

I loaded the different package in RStudio such as: broom, tidyverse, gtsummary, tableone and loaded by two different datasets into RStudio:

```
## load file
````{r}
Conditionscontributingcovid19USA <-
read.csv("C:/users/shivk/downloads/ConditionscontributingCovid19USA.csv")
summary(ConditionscontributingCovid19USA)
states <- read.csv("C:/users/shivk/downloads/states.csv")
summary(states)
```

I removed noisy variables ICD10\_codes, Flag, Date as of, start week, end week from our final dataset because these variables act as constant and they are not contributing anything to the predicted response variable.

```

## remove unwanted variable
```{r}
ConditionscontributingCovid19USA$Data.as.of <- NULL
ConditionscontributingCovid19USA$Start.Week <- NULL
ConditionscontributingCovid19USA$End.Week <- NULL
ConditionscontributingCovid19USA$ICD10_codes <- NULL
ConditionscontributingCovid19USA$Flag <- NULL

```

We have all the independent variables as categorical variables as showing “char” so in order to create a model, I created dummy variables for each of the independent categorical variables. I joined the states dataset by state code into coviddataUSA\_clean datasets. I relevel the dummy variables that I want to keep as reference for other dummy variables.

```

## data cleaning
```{r}
coviddataUSA_clean <- covid_select %>%
  filter(state != 'us') %>%
  filter( age_group != ' Not stated') %>%
  filter( age_group != ' All ages') %>%
  filter( age_group %in% c('0-24', '25-34')) %>%
  inner_join(states, by=c('state' = 'state_code')) %>%
  
  transmute(condition_group = as.factor(condition_group)
           , condition = as.factor(condition)
           , age_group = as.factor(age_group)
           , region = as.factor(region)
           , num_deaths = num_deaths) %>%
  mutate(condition_group = relevel(condition_group, ref ='Respiratory diseases')
         , condition = relevel(condition, ref ='Respiratory failure')
         , age_group = relevel(age_group, ref ='0-24')
         , region = relevel(region, ref ='Midwest')) %>%
  na.omit()

## create matrix for agegroup
agegroupdummies.matrix <- model.matrix(~ coviddataUSA_clean$age_group)
## convert the model matrix into a data frame
agegroupdummies.frame <- data.frame(agegroupdummies.matrix)
coviddataUSA_clean <- cbind(coviddataUSA_clean, agegroupdummies.frame)

## create matrix for condition
conditiondummies.matrix <- model.matrix(~ coviddataUSA_clean$condition)
## convert the model matrix into a data frame
conditiondummies.frame <- data.frame(conditiondummies.matrix)
coviddataUSA_clean <- cbind(coviddataUSA_clean, conditiondummies.frame)

## create matrix for conditiongroup
conditiongroupdummies.matrix <- model.matrix(~ coviddataUSA_clean$condition_group)
## convert the model matrix into a data frame
conditiongroupdummies.frame <- data.frame(conditiongroupdummies.matrix)
coviddataUSA_clean <- cbind(coviddataUSA_clean, conditiongroupdummies.frame)

## create matrix for regions
regiondummies.matrix <- model.matrix(~ coviddataUSA_clean$region)
## convert the model matrix into a data frame
regiondummies.frame <- data.frame(regiondummies.matrix)
coviddataUSA_clean <- cbind(coviddataUSA_clean, regiondummies.frame)

```

```

## select the variable of interest after data cleaning
``{r}
covid19dataUSA <- coviddatousa_clean %>%
  select( num_deaths = num_deaths
         , age25_34 = coviddatousa_clean.age_group25_34
         , allothercondandcauses =
coviddatousa_clean.condition_groupAll.other.conditions.and.causes..residual.
         , alzheimer = coviddatousa_clean.condition_groupAlzheimer.disease
         , circulatorydiseases = coviddatousa_clean.condition_groupCirculatory.diseases
         , covid19 = coviddatousa_clean.condition_groupCOVID.19
         , diabetes = coviddatousa_clean.condition_groupDiabetes
         , injuryPoisoning =
coviddatousa_clean.condition_groupIntentional.and.unintentional.injury..poisoning..and.other.adverse.events
         , malignant = coviddatousa_clean.condition_groupMalignant.neoplasms
         , obesity = coviddatousa_clean.condition_groupobesity
         , renalFailure = coviddatousa_clean.condition_groupRenal.failure
         , sepsis = coviddatousa_clean.condition_groupSepsis
         , vascular = coviddatousa_clean.condition_groupVascular.and.unspecified.dementia
         , south = coviddatousa_clean.regionSouth
         , west = coviddatousa_clean.regionWest
         , northeast = coviddatousa_clean.regionNortheast)
str(covid19dataUSA)
```

```

Before executing above code, I built the model with same covid19 datasets with different name and checked for multicollinearity.

In the first milestone I focused especially on the “age dataset” and how these age groups 0-24 and 25-34 along with conditions and condition groups excluding states are contributing to increase the total number of COVID19 deaths in the USA. I created regression model as shown below using regression equation:

```
> Model1 <- lm(NoOfCovid19Deaths ~ Age.Group + Condition.Group + Condition, data = ConditionscontributingCovid19USA.pinki)
```

```

> summary(Model1)

Call:
lm(formula = NoOfCovid19Deaths ~ Age.Group + Condition.Group +
   Condition, data = ConditionscontributingCovid19USA.pinki)

Residuals:
    Min      1Q  Median      3Q     Max 
 -40.61   -6.53   -2.11    1.23 1446.39 

Coefficients: (11 not defined because of singularities)
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         20.4158   4.1849  4.878  1.14e-06 ***
Age.Group25-34                      7.0199   1.7171  4.088  4.49e-05 ***
Condition.GroupAlzheimer disease    -23.9257   5.7924 -4.131  3.74e-05 ***
Condition.GroupCirculatory diseases -18.7812   5.8197 -3.227  0.001267 ** 
Condition.GroupCOVID-19              13.1726   5.7924  2.274  0.023046 *  
Condition.GroupDiabetes             -16.9076   5.7924 -2.919  0.003545 ** 
Condition.GroupIntentional and unintentional injury, poisoning, and other adverse events -18.3095   5.8478 -3.131  0.001763 ** 
Condition.GroupMalignant neoplasms   -20.8479   5.7924 -3.599  0.000326 *** 
Condition.GroupObesity              -13.2477   5.7924 -2.287  0.022277 *  
Condition.GroupRenal failure        -18.6694   5.7924 -3.223  0.001285 ** 
Condition.GroupRespiratory diseases -14.9453   5.7924 -2.580  0.009934 ** 
Condition.GroupSepsis               -17.3925   5.7924 -3.003  0.002704 ** 
Condition.GroupVascular and unspecified dementia -23.9257   5.8197 -4.111  4.07e-05 *** 
conditionAll other conditions and causes (residual) NA         NA      NA      NA      
ConditionAlzheimer disease          NA         NA      NA      NA      
ConditionCardiac arrest             2.5478   5.8748  0.434  0.664560    
ConditionCardiac arrhythmia        -2.7268   5.8197 -0.469  0.639434    
ConditionCerebrovascular diseases  -2.8634   5.8468 -0.490  0.624359    
ConditionChronic lower respiratory diseases -5.4026   5.7924 -0.933  0.351067    
ConditionCOVID-19                  NA         NA      NA      NA      
ConditionDiabetes                 -2.6157   5.8197 -0.449  0.653142    
ConditionHeart failure             -0.7365   5.8197 -0.127  0.899307    
ConditionHypertensive diseases    8.5097   5.7924  1.469  0.141932    
ConditionInfluenza and pneumonia  NA         NA      NA      NA      
ConditionIntentional and unintentional injury, poisoning, and other adverse events NA         NA      NA      NA      
ConditionIschemic heart disease   -3.6426   5.8197 -0.626  0.531428    
ConditionMalignant neoplasms       NA         NA      NA      NA      
ConditionObesity                  NA         NA      NA      NA      
ConditionOther diseases of the circulatory system NA         NA      NA      NA      
ConditionOther diseases of the respiratory system -5.2887   5.7924 -0.913  0.361309    
ConditionRenal failure             NA         NA      NA      NA      
ConditionRespiratory arrest       -6.4999   5.9707 -1.089  0.276425    
ConditionRespiratory failure      4.8426   5.7924  0.836  0.403223    
ConditionSepsis                   NA         NA      NA      NA      
ConditionVascular and unspecified dementia NA         NA      NA      NA      
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.57 on 2434 degrees of freedom
Multiple R-squared:  0.04444, Adjusted R-squared:  0.03541 
F-statistic: 4.922 on 23 and 2434 DF,  p-value: 1.601e-13

```

Here, in the above regression model, when we look at the F test, p- value is statistically significant as its value is less than alpha level (0.05). So, we reject the null hypothesis that all the betas are equal to zero and accept alternative hypothesis that at least one of the betas is not equal to zero. Similarly, when we look at the T test, p- value for each of the dummy variables, we see all the dummy variables of the Condition group and age group are statistically significant. That means p- value is less than alpha level (0.05) so we reject the null hypothesis that beta 1 to beta 12 are equal to zero and accept the alternative hypothesis that beta 1 to beta12 are not equal to zero.

On the other hand, when we looked at the T test, p- value of all the dummy variables of condition are insignificant. There are 11 NA's (Not Applicable) which are not defined because of singularities.

The adjusted R Square is 3.541% this means our model predicts 3.541% variability in dependent variables. This is extremely low. By looking at the model we assume either there is something missing or there is multicollinearity in our dataset. To detect it we use Variance inflation factor formula:

```
> vif(Model1)
Error in vif.default(Model1) :
  there are aliased coefficients in the model
```

There are aliased coefficients in the model that means variables are linearly dependent on others (i.e., cause perfect multicollinearity) so need identify which variables are culprit in our model using a formula:

```
> alias(Model1)
```

| Complete :                                                                         | (Intercept) | Age.Group25-34 | Condition.GroupAlzheimer disease |
|------------------------------------------------------------------------------------|-------------|----------------|----------------------------------|
| ConditionAll other conditions and causes (residual)                                | 1           | 0              | -1                               |
| ConditionAlzheimer disease                                                         | 0           | 0              | 1                                |
| ConditionCOVID-19                                                                  | 0           | 0              | 0                                |
| ConditionDiabetes                                                                  | 0           | 0              | 0                                |
| ConditionIntentional and unintentional injury, poisoning, and other adverse events | 0           | 0              | 0                                |
| ConditionMalignant neoplasms                                                       | 0           | 0              | 0                                |
| ConditionObesity                                                                   | 0           | 0              | 0                                |
| ConditionOther diseases of the circulatory system                                  | 0           | 0              | 0                                |
| ConditionRenal failure                                                             | 0           | 0              | 0                                |
| ConditionSepsis                                                                    | 0           | 0              | 0                                |
| ConditionVascular and unspecified dementia                                         | 0           | 0              | 0                                |
| ConditionAll other conditions and causes (residual)                                | -1          | -1             | -1                               |
| ConditionAlzheimer disease                                                         | 0           | 0              | 0                                |
| ConditionCOVID-19                                                                  | 0           | 1              | 0                                |
| ConditionDiabetes                                                                  | 0           | 0              | 1                                |
| ConditionIntentional and unintentional injury, poisoning, and other adverse events | 0           | 0              | 0                                |
| ConditionMalignant neoplasms                                                       | 0           | 0              | 0                                |
| ConditionObesity                                                                   | 0           | 0              | 0                                |
| ConditionOther diseases of the circulatory system                                  | 1           | 0              | 0                                |
| ConditionRenal failure                                                             | 0           | 0              | 0                                |
| ConditionSepsis                                                                    | 0           | 0              | 0                                |
| ConditionVascular and unspecified dementia                                         | 0           | 0              | 0                                |
| Condition.GroupCirculatory diseases                                                | 0           | 0              | 0                                |
| Condition.GroupCOVID-19                                                            | 0           | 0              | 0                                |
| Condition.GroupDiabetes                                                            | 0           | 0              | 0                                |

The alias table shows that all those terms are linearly dependent on all the dummy variables of Condition. This means they are highly correlated. So, I removed it and build the model and did cross validation to validate the model. Finally got first order model without states found adjusted R-square still too low.

```
> Model2 <- lm(NoOfCovid19Deaths ~ Age.Group + Condition.Group, data =
  ConditionscontributingCovid19USA.pinki)
```

```
> summary(Model2)

Call:
lm(formula = NoOfCovid19Deaths ~ Age.Group + Condition.Group,
    data = ConditionscontributingCovid19USA.pinki)

Residuals:
    Min      1Q   Median     3Q    Max 
-40.61  -6.59  -2.74   0.46 1446.39 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.416   4.187  4.877 1.15e-06 ***
Age.Group25-34          7.020   1.718  4.087 4.52e-05 ***
Condition.GroupAlzheimer disease -23.926   5.795 -4.129 3.77e-05 ***
Condition.GroupCirculatory diseases -20.236   4.383 -4.617 4.10e-06 ***
Condition.GroupCOVID-19           13.173   5.795  2.273 0.023099 *  
Condition.GroupDiabetes         -16.908   5.795 -2.918 0.003557 ** 
Condition.GroupIntentional and unintentional injury, poisoning, and other adverse events -18.309   5.850 -3.138 0.001770 ** 
Condition.GroupMalignant neoplasms -20.848   5.795 -3.598 0.000327 *** 
Condition.GroupObesity          -13.248   5.795 -2.286 0.022329 *  
Condition.GroupRenal failure     -18.669   5.795 -3.222 0.001291 ** 
Condition.GroupRespiratory diseases -15.475   4.432 -3.492 0.000488 *** 
Condition.GroupSepsis            -17.392   5.795 -3.001 0.002714 ** 
Condition.GroupVascular and unspecified dementia -23.926   5.822 -4.110 4.09e-05 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.58 on 2445 degrees of freedom
Multiple R-squared:  0.03937, Adjusted R-squared:  0.03465 
F-statistic: 8.35 on 12 and 2445 DF,  p-value: 1.143e-15
```

In above model I found the age group 25-34 and condition group alzheimer disease and covid19 itself contributing to the increase in number of covid19 deaths. Our model only explained 3.465% variability in response variable. The f-test looks good that we rejected the null hypothesis all betas are equal to zero and accept the alternative that at least one of the betas is not equal to zero. Looking at the t-test, all the dummy variables are significant.

In this milestone2, I focused especially on the “age dataset” and how these age groups 0-24 and 25-34 along with conditions and region are contributing to increase the total number of COVID19 deaths in the USA. I created a first order model including new independent variables created by grouping 51 states into 4 regions, along with age group, conditions group and removing conditions group as it was causing multicollinearity.

```
> fullModel <- lm(num_deaths ~ age25_34 + allothercondandcauses + alzheimer + ci
rculatorydiseases + covid19 + diabetes + injuryPoisoning + malignant + obesity +
renalFailure + sepsis + vascular + south + west + northeast, data = covid19data
USA )
> summary(fullModel)

Call:
lm(formula = num_deaths ~ age25_34 + allothercondandcauses +
alzheimer + circulatorydiseases + covid19 + diabetes + injuryPoisoning +
malignant + obesity + renalFailure + sepsis + vascular +
south + west + northeast, data = covid19dataUSA)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.772 -2.888 -0.922  1.213 160.016 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.2662    0.7520   0.354  0.723421    
age25_34     4.1009    0.5329   7.696 2.56e-14 ***  
allothercondandcauses 12.8459   1.4596   8.801 < 2e-16 ***  
alzheimer    -3.5845   1.1492  -3.119 0.001850 **  
circulatorydiseases -2.7511   0.7174  -3.835 0.000131 ***  
covid19       24.1332   1.5055  16.030 < 2e-16 ***  
diabetes      -1.3271   1.5026  -0.883 0.377256    
injuryPoisoning -1.7362   1.3354  -1.300 0.193747    
malignant     -3.6101   1.3368  -2.701 0.006999 **  
obesity       2.5789   1.5991   1.613 0.107017    
renalFailure  -1.8813   1.4145  -1.330 0.183740    
sepsis        -1.6485   1.5410  -1.070 0.284920    
vascular      -3.5844   1.1582  -3.095 0.002006 **  
south         2.4840    0.7213   3.444 0.000589 ***  
west          1.2715    0.7369   1.725 0.084648 .  
northeast     0.6559    0.8204   0.799 0.424136    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.17 on 1478 degrees of freedom
Multiple R-squared:  0.276,    Adjusted R-squared:  0.2686 
F-statistic: 37.55 on 15 and 1478 DF,  p-value: < 2.2e-16
```

We observed when we added region and built the model, adjusted R-square value increased which was 26.86%. This means our model explained 26.86% variability in dependent variables. The F-test p value it was statistically significant and however looking at the T- test some of the variables were not significant and fail to reject the null hypothesis that beta associated to the variable equal to zero. Finally, I thought of trying feature selections to find the variables which are the best predictor for predicting response variable. First, I ran

backward elimination with the full model by removing variables one by one at a time. Final model remained with 12 independent variables is shown below:

```

library(MASS)
Model_full <- lm(num_deaths ~., data = covid19dataUSA)
step <- stepAIC(Model_full, direction = "backward")
step$anova
summary(step)
```
> Model_finalBakward <- lm(num_deaths ~ age25_34 + allothercondandcauses + alzheimer + circulatorydiseases + covid19 + malignant + obesity + vascular + south + west, data = covid19dataUSA)
> summary(Model_finalBakward)

Call:
lm(formula = num_deaths ~ age25_34 + allothercondandcauses +
alzheimer + circulatorydiseases + covid19 + malignant + obesity +
vascular + south + west, data = covid19dataUSA)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.816  -3.234  -0.857   1.226 159.971 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.1328    0.5654  -0.235 0.814345    
age25_34      4.1187    0.5324   7.737 1.87e-14 ***  
allothercondandcauses 13.5341   1.4163   9.556 < 2e-16 ***  
alzheimer     -2.9127   1.0940  -2.662 0.007841 **  
circulatorydiseases -2.0830   0.6252  -3.332 0.000884 ***  
covid19        24.8409   1.4632  16.977 < 2e-16 ***  
malignant       -2.9541   1.2892  -2.291 0.022083 *  
obesity         3.2579   1.5597   2.089 0.036896 *  
vascular        -2.9126   1.1034  -2.640 0.008385 **  
south           2.2016   0.6277   3.508 0.000466 ***  
west            0.9895   0.6455   1.533 0.125524  
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.17 on 1483 degrees of freedom
Multiple R-squared:  0.2738,    Adjusted R-squared:  0.2689
F-statistic: 55.91 on 10 and 1483 DF,  p-value: < 2.2e-16

```

I ran the forward selection method with the empty model.

```

library (MASS)
model_full <- lm(num_deaths ~., data = covid19dataUSA)
model_empty <- lm(num_deaths ~ 1, data = covid19dataUSA)
step <- stepAIC(model_empty, direction = "forward", scope = list(upper = model_full, lower=
model_empty))
step$anova
summary(step)

```

```

> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
num_deaths ~ 1

Final Model:
num_deaths ~ covid19 + allothercondandcauses + age25_34 + obesity +
  south + circulatorydiseases + alzheimer + vascular + malignant +
  west

      Step Df  Deviance Resid. Df Resid. Dev      AIC
1              1493  211249.3 7399.664
2 + covid19  1 34800.8195 1492  176448.5 7132.729
3 + allothercondandcauses 1 13192.5633 1491  163256.0 7018.630
4 + age25_34 1 5782.1029 1490  157473.9 6966.757
5 + obesity   1  947.6776 1489  156526.2 6959.739
6 + south     1  894.4503 1488  155631.7 6953.177
7 + circulatorydiseases 1  401.9787 1487  155229.7 6951.313
8 + alzheimer 1  450.6256 1486  154779.1 6948.970
9 + vascular  1  587.9420 1485  154191.2 6945.284
10 + malignant 1  534.1225 1484  153657.1 6942.100
11 + west     1  243.0684 1483  153414.0 6941.734

```

Call:

```

lm(formula = num_deaths ~ covid19 + allothercondandcauses + age25_34 +
  obesity + south + circulatorydiseases + alzheimer + vascular +
  malignant + west, data = covid19dataUSA)

```

Residuals:

| Min     | 1Q     | Median | 3Q    | Max     |
|---------|--------|--------|-------|---------|
| -29.816 | -3.234 | -0.857 | 1.226 | 159.971 |

Coefficients:

|                       | Estimate | Std. Error | t value | Pr(> t )     |
|-----------------------|----------|------------|---------|--------------|
| (Intercept)           | -0.1328  | 0.5654     | -0.235  | 0.814345     |
| covid19               | 24.8409  | 1.4632     | 16.977  | < 2e-16 ***  |
| allothercondandcauses | 13.5341  | 1.4163     | 9.556   | < 2e-16 ***  |
| age25_34              | 4.1187   | 0.5324     | 7.737   | 1.87e-14 *** |
| obesity               | 3.2579   | 1.5597     | 2.089   | 0.036896 *   |
| south                 | 2.2016   | 0.6277     | 3.508   | 0.000466 *** |
| circulatorydiseases   | -2.0830  | 0.6252     | -3.332  | 0.000884 *** |
| alzheimer             | -2.9127  | 1.0940     | -2.662  | 0.007841 **  |
| vascular              | -2.9126  | 1.1034     | -2.640  | 0.008385 **  |
| malignant             | -2.9541  | 1.2892     | -2.291  | 0.022083 *   |
| west                  | 0.9895   | 0.6455     | 1.533   | 0.125524     |

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.17 on 1483 degrees of freedom  
 Multiple R-squared: 0.2738, Adjusted R-squared: 0.2689  
 F-statistic: 55.91 on 10 and 1483 DF, p-value: < 2.2e-16

From our final first order model, I removed the west as it was insignificant and even after removing the dummy variable west, there was not that difference in adjusted R-square value. The difference was only 0.0007

```

> Model_finalForward <- lm(num_deaths ~ covid19 + allothercondandcauses + age25_34 + obesity + south + circulatorydiseases + alzheimer + vascular + malignant, data = covid19dataUSA)
> summary(Model_finalForward)

Call:
lm(formula = num_deaths ~ covid19 + allothercondandcauses + age25_34 +
    obesity + south + circulatorydiseases + alzheimer + vascular +
    malignant, data = covid19dataUSA)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.206 -3.265 -0.248  1.827 159.983 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.2478   0.5082  0.488  0.62586    
covid19       24.8222  1.4638 16.957 < 2e-16 ***
allothercondandcauses 13.5304  1.4170  9.549 < 2e-16 ***
age25_34      4.1355  0.5325  7.766  1.5e-14 ***
obesity        3.3048  1.5601  2.118  0.03431 *  
south          1.8112  0.5739  3.156  0.00163 ** 
circulatorydiseases -2.0749  0.6255 -3.317  0.00093 *** 
alzheimer     -2.9193  1.0945 -2.667  0.00773 ** 
vascular       -2.9314  1.1038 -2.656  0.00800 ** 
malignant      -2.9292  1.2897 -2.271  0.02328 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.18 on 1484 degrees of freedom
Multiple R-squared:  0.2726, Adjusted R-squared:  0.2682 
F-statistic: 61.8 on 9 and 1484 DF, p-value: < 2.2e-16

```

In order to validate model, ModelA, the final model that I got from backward elimination method, ModelB, the final model that I got from forward elimination and ModelC with all the variables I used n fold cross validation process. I created a training and testing set using n fold cross validation, with an 80-20 split to validate. I have a dataset to train (1202obs.) and to test (292obs.).

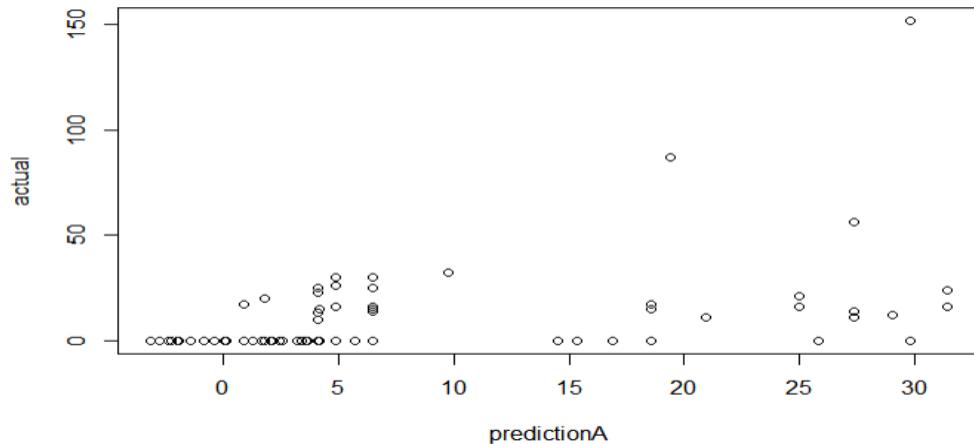
```

set.seed(123)
partition <- sample(2,nrow(covid19dataUSA), replace = TRUE, prob = c(0.80, 0.20))
train <- covid19dataUSA[partition == 1 ,]
test <- covid19dataUSA[partition == 2 ,]

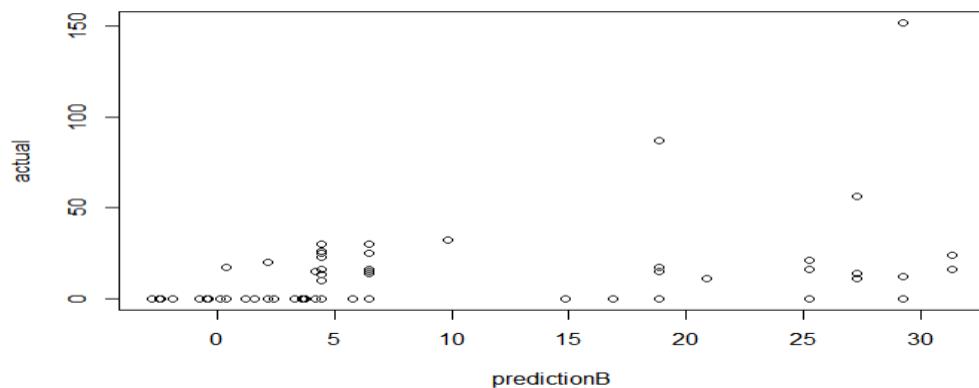
```
> predictionA <- predict(ModelA, test)
> actual = test$num_deaths
> cor(predictionA, actual)
[1] 0.485
> plot(predictionA, actual)
> predictionB <- predict(ModelB, test)
> actual = test$num_deaths
> cor(predictionB, actual)
[1] 0.482
> plot(predictionB, actual)
>
> predictionC <- predict(ModelC, test)
> actual = test$num_deaths
> cor(predictionC, actual)
[1] 0.484
> plot(predictionC, actual)
```

```

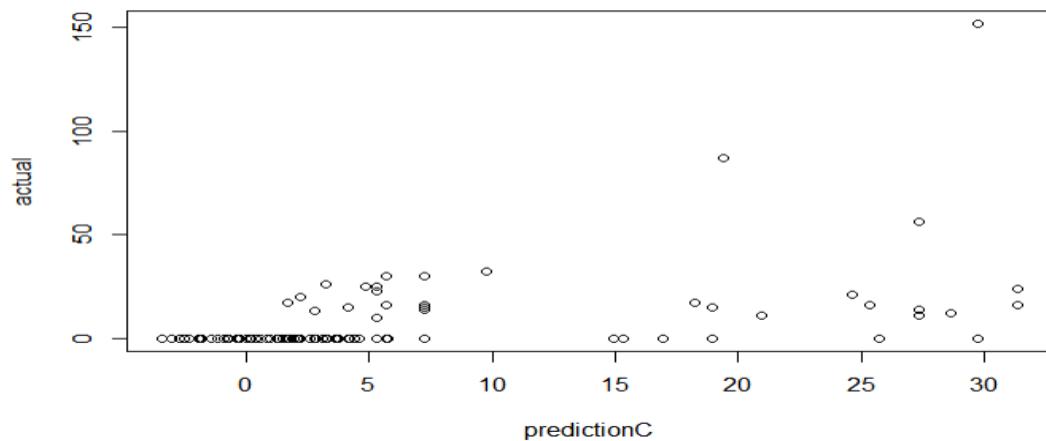
The ModelA with the train data have adjusted R-square =27.79%



The ModelB with the train data have adjusted R-square =27.78%



The ModelC with the train data have adjusted R-square =27.78%

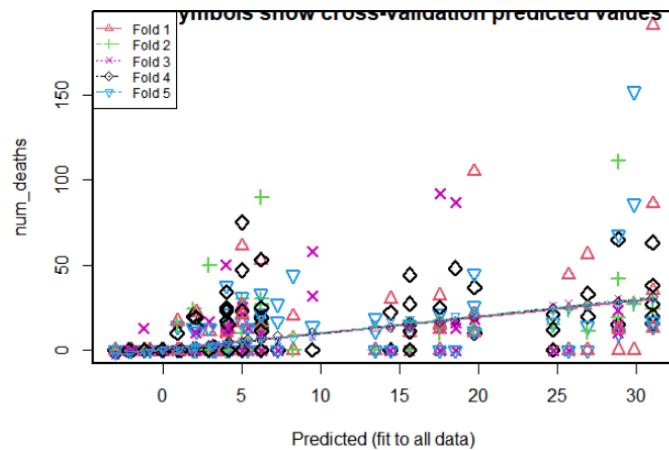


The model created from training dataset is useful for prediction. The correlation between actual and prediction is 0.485 for ModelA and 0.482 for ModelB and 0.484 for ModelC. As

we look at the figure above, scatter plots of actual versus prediction, data points are spread around but there are lots of outliers.

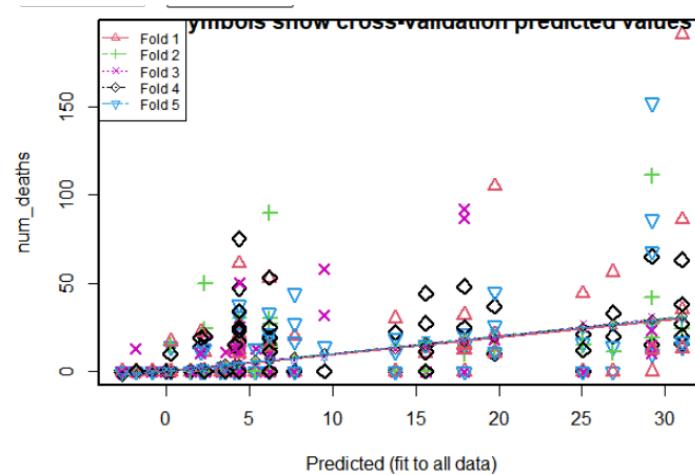
The goal of Cross validation is to test the model's ability to predict new data which was not used in estimating it. Depending on data size, I created n-fold cross validation with 5-folds.

```
> outA <- cv.lm(data = covid19dataUSA, form.lm = formula(num_deaths ~ age25_34 + allothe
rcondandcauses + alzheimer + circulatorydiseases + covid19 + malignant + obesity + vascu
lar + south + west), plotit = "Observed", m=5)
```



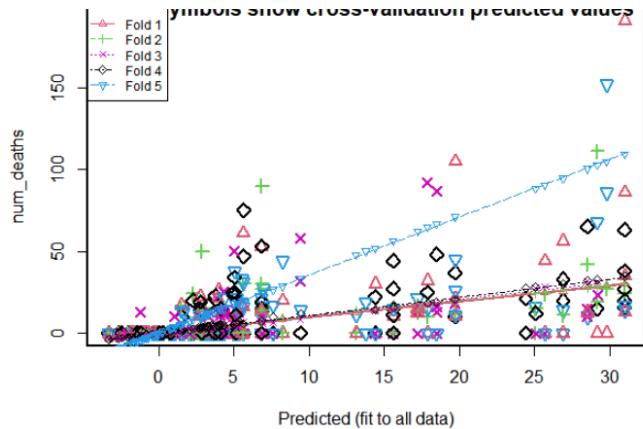
Here, in above figure it is clearly, seen 3,5-fold held straight line in a plot. I got an average mean square 150.

```
> outB <- cv.lm(data = covid19dataUSA, form.lm = formula(num_deaths ~ covid19 + allother
condandcauses + age25_34 + obesity + south + circulatorydiseases + alzheimer + vascular
+ malignant), plotit = "Observed", m=5)
```



After doing n fold cross validation for ModelB got average mean square value 150 and observed fold 3 and 5 held straight line in the plot.

```
> outC <- cv.lm(data= covid19dataUSA, form.lm = formula(num_deaths ~.), plotit = "Observ
ed", m=5)
```



After doing n fold cross validation for ModelC, the average mean square is 170 and observed that fold 5 held a straight line in the plot and fold 1 and 4 also held another straight line. I already tried to detect multicollinearity, using a variance inflation factor found independent variable condition is causing multicollinearity. I already removed condition variables and create the model.

Comparing adjusted, correlation value and average mean square of all the models, the first order model will be ModelB (forward selection).

```

Model_finalForward <- lm(num_deaths ~ covid19 + allothercondandcauses + age25_34 + obesity +
south + circulatorydiseases + alzheimer + vascular + malignant, data = covid19dataUSA)
summary(Model_finalForward)

> Model_finalForward <- lm(num_deaths ~ covid19 + allothercondandcauses + age25_34 + obe
sity + south + circulatorydiseases + alzheimer + vascular + malignant, data = covid19dat
USA)
> summary(Model_finalForward)

Call:
lm(formula = num_deaths ~ covid19 + allothercondandcauses + age25_34 +
obesity + south + circulatorydiseases + alzheimer + vascular +
malignant, data = covid19dataUSA)

Residuals:
    Min      1Q      Median      3Q      Max 
-29.206 -3.265 -0.248  1.827 159.983 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.2478     0.5082   0.488   0.62586    
covid19     24.8222    1.4638  16.957 < 2e-16 ***  
allothercondandcauses 13.5304    1.4170   9.549 < 2e-16 ***  
age25_34    4.1355     0.5325   7.766  1.5e-14 ***  
obesity     3.3048     1.5601   2.118   0.03431 *    
south       1.8112     0.5739   3.156   0.00163 **  
circulatorydiseases -2.0749    0.6255  -3.317   0.00093 ***  
alzheimer   -2.9193    1.0945  -2.667   0.00773 **  
vascular    -2.9314    1.1038  -2.656   0.00800 **  
malignant   -2.9292    1.2897  -2.271   0.02328 *    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.18 on 1484 degrees of freedom
Multiple R-squared:  0.2726,    Adjusted R-squared:  0.2682 
F-statistic: 61.8 on 9 and 1484 DF,  p-value: < 2.2e-16

```

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 and

obesity 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 COVID19 deaths.

After building best fit first order I tried finding best interaction term using subset selection and forward and backward feature selection. Finally got these below interaction terms as best using subsets selection method prune insignificant interaction one by one at a time because with this interaction there was a slight increase in adjusted R- square even though some of the value is insignificant.

```
> model.interactionA6 <- lm(num_deaths ~ covid19 + allothercondandcauses +
+   age25_34 + obesity + south + circulatorydiseases + alzheimer + vascular +
+   malignant + covid19*age25_34 + allothercondandcauses*age25_34 + obesity*age25_34 + circulatorydiseases*age25_34 + alzheimer*age25_34 + vascular*age25_34 + malignant*age25_34 + covid19*south + allothercondandcauses*south, data = covid19dataUSA)
>
> summary(model.interactionA6)

call:
lm(formula = num_deaths ~ covid19 + allothercondandcauses + age25_34 +
  obesity + south + circulatorydiseases + alzheimer + vascular +
  malignant + covid19 * age25_34 + allothercondandcauses *
  age25_34 + obesity * age25_34 + circulatorydiseases * age25_34 +
  alzheimer * age25_34 + vascular * age25_34 + malignant *
  age25_34 + covid19 * south + allothercondandcauses * south,
  data = covid19dataUSA)

Residuals:
    Min      1Q  Median      3Q     Max 
-36.926 -1.262 -0.712  0.340 145.389 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.07202  0.56138  0.128  0.897932    
covid19       8.95039  2.32173  3.855  0.000121 ***  
allothercondandcauses  6.11048  2.35880  2.591  0.009678 **  
age25_34      4.94529  0.82311  6.008  2.36e-09 ***  
obesity        0.77600  1.83470  0.423  0.672389    
south          1.19016  0.57504  2.070  0.038654 *    
circulatorydiseases -0.41179  0.79144 -0.520  0.602927    
alzheimer      -0.46874  1.46995 -0.319  0.749861    
vascular        -0.47668  1.48271 -0.321  0.747884    
malignant       -0.50180  1.71498 -0.293  0.769870    
covid19:age25_34 22.95807  2.83255  8.105  1.10e-15 ***  
allothercondandcauses:age25_34  8.80152  2.76388  3.184  0.001480 **  
age25_34:obesity      8.02265  3.17982  2.523  0.011741 *    
age25_34:circulatorydiseases -3.89368  1.21440 -3.206  0.001374 **  
age25_34:alzheimer     -4.94529  2.10171 -2.353  0.018754 *    
age25_34:vascular      -4.94529  2.11943 -2.333  0.019766 *    
age25_34:malignant     -4.97057  2.47617 -2.007  0.044893 *    
covid19:south         7.49542  2.83026  2.648  0.008176 **  
allothercondandcauses:south  5.67678  2.74438  2.069  0.038766 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.765 on 1475 degrees of freedom
Multiple R-squared:  0.3342, Adjusted R-squared:  0.326 
F-statistic: 41.12 on 18 and 1475 DF,  p-value: < 2.2e-16
```

Looking at the F test, p –value was less than alpha level (0.05), we rejected null hypothesis that all the betas are equal to zero and accepted the alternative hypothesis that at least one of the betas is not equal to zero. The adjusted R-square is 32.6% which means our interaction model predict 32.6% variability in number of covid19 deaths (dependent variable). Looking at the T-test, some of the dummy variables p-value is insignificant means fail to reject null hypothesis. All the interaction terms are significant; we need to include its children even though some of the independent variables were insignificant.

In above summary, we found people of the age group 25\_34 leaving in the south region of the USA suffered from condition group like all other diseases and causes, obesity and

covid19 itself are contributing a greater number of covid19 deaths when compared with age group 0-24 and region midwest and condition group respiratory diseases.

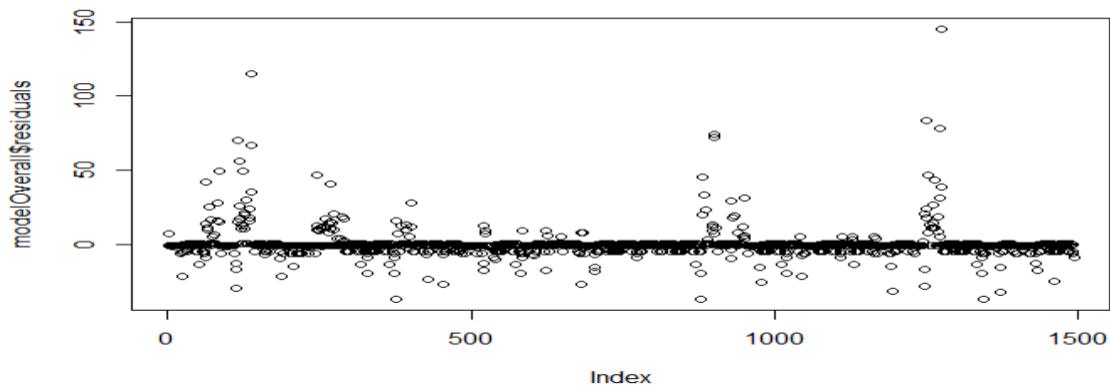
Then using interaction model, I evaluated four assumptions of the residuals and did residuals analysis

```
modelOverall <- lm(num_deaths ~ covid19 + allothercondandcauses +
  age25_34 + obesity + south + circulatorydiseases + alzheimer + vascular +
  malignant + covid19*age25_34 + allothercondandcauses*age25_34 + obesity*age25_34 +
  circulatorydiseases*age25_34 + alzheimer*age25_34 + vascular*age25_34 + malignant*age25_34 +
  covid19*south + allothercondandcauses*south, data = covid19dataUSA)
plot(modelOverall$residuals)
```

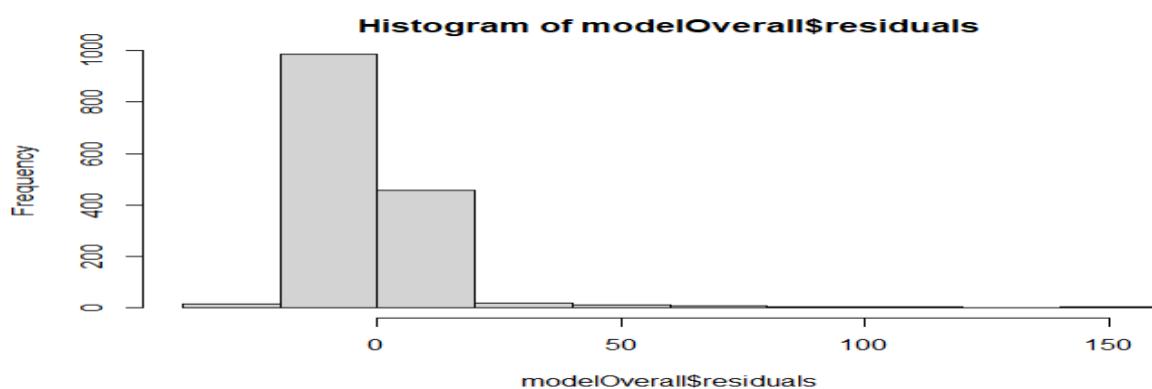
I verified the assumption that the sum of the residuals should add up to zero by using sum()

```
> sum(modelOverall$residuals)
[1] 3.04376e-12
```

The sum of the residuals almost equal to zero we can assume our model passed this assumption



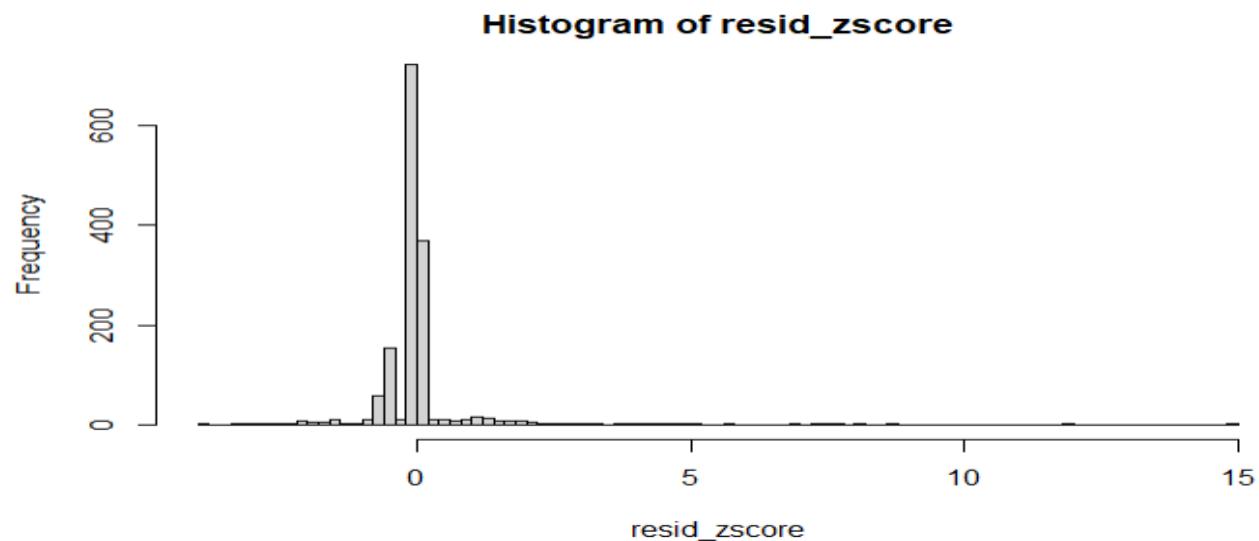
The residuals have constant variance.



In the above histogram of model residuals, we observed data points are right skewed and there is an outlier on the right side.

Second assumption residuals are normally distributed to verified I calculate mean, standard deviation and resid\_zscore and finally plot the histogram

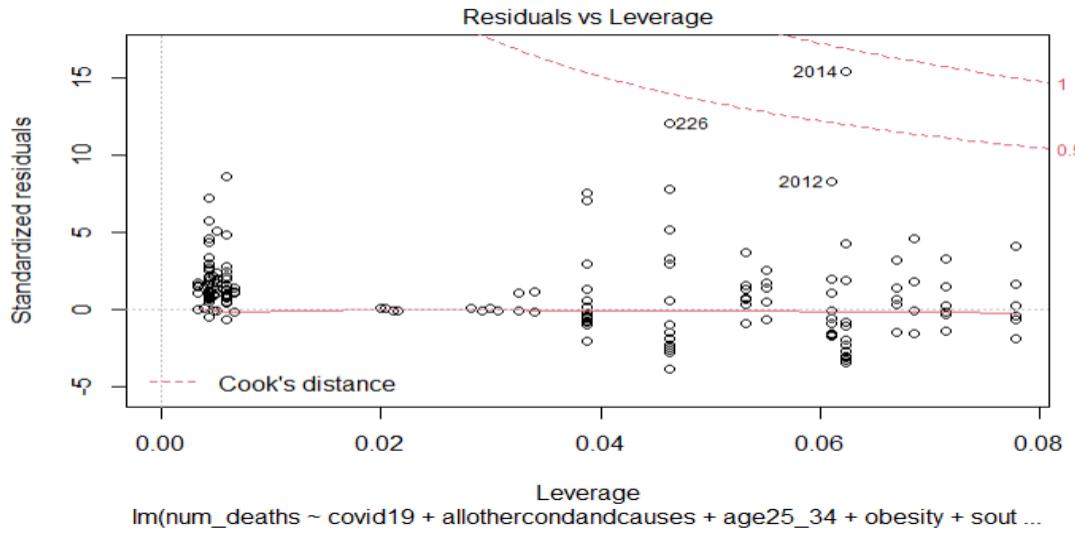
```
mean = mean(modeloverall$residuals) #
sd = sd(modeloverall$residuals)
resid_zscore = (modeloverall$residuals - mean)/sd
hist(resid_zscore, breaks = 100)
````
```



In the above histogram, we assume that 95 percent of the residuals are within two standard deviations and there are outliers on the right and left side something we need to be aware of. I used the Durbin Watson test to verify the assumption that the residuals should be independent:

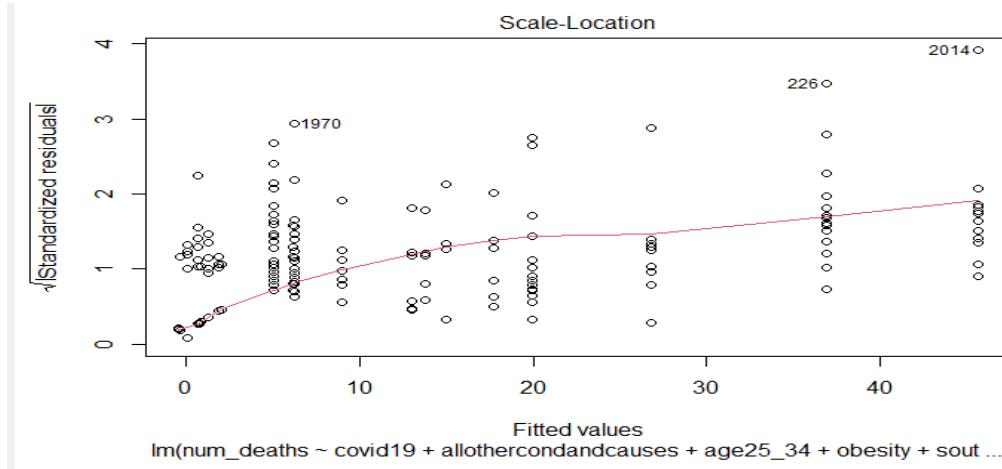
```
lag Autocorrelation D-W Statistic p-value
 1      0.3627491    1.273912      0
Alternative hypothesis: rho != 0
```

Since the computed D-W statistic falls below the lower value of this range, we reject the null hypothesis that residuals are not correlated. Hence, an alternate hypothesis that there may be a correlation between residuals comes into play.

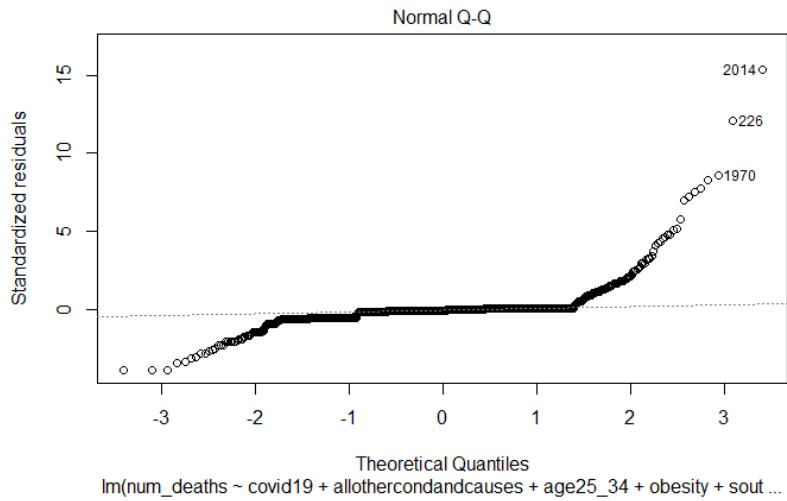


This plot is used to find influential cases in the dataset. An influential case is one that, if removed, will affect the model so its inclusion or exclusion should be considered.

In the above residuals versus leverage plot, there are few leverage points lies far from the trend which we need handle. The plot for covid19dataUSA example shows threes cases are outside Cooks Distance.

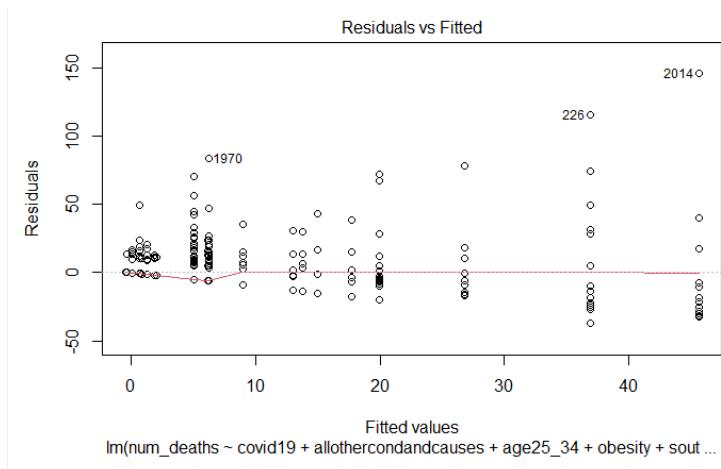


This plot tests the linear regression assumption of equal variance (homoscedasticity) i.e., that the residuals have equal variance along the regression line. It is also called the Spread-Location plot. Also, called fitted value against the square root of standardized residuals plot, we observed constant variance on the right side of the plot and few of the outliers in our dataset. However, the beginning of the line does have fewer points so slightly less variance.



Residuals are normally distributed, and the Q-Q Plot will show this. If residuals follow close to a straight line on this plot, it is a good indication they are normally distributed.

When we looked at the residual QQ plot, clearly illustrate our dataset is normal as all the datapoints are line up on the straight diagonal line. However, there are some outliers in our dataset which we observed, and some deviation is there on right side and left side.



Looking at the above residual plot, there is a multiplicative pattern in the variance especially as we go through left to right so instead of predicting the number of deaths, we used log transformation for dependent variable; the log of num\_death plus 1, that is to say,  $\log(\text{num\_death} + 1)$

The best final model that I got after removing the outliers.

```

> model.interactionH <- lm(log(num_deaths + 1) ~ covid19 + allothercondandcauses +
+   age25_34 + obesity + south + circulatorydiseases + alzheimer + vascular +
+   malignant + covid19*age25_34 + allothercondandcauses*age25_34 + obesity*age25_34 + circulatorydiseases*age25_34 + alzheimer*age25_34 + vascular*age25_34 + malignant*age25_34 + covid19*south + allothercondandcauses*south, data = covid19dataUSA)
>
> summary(model.interactionH)

Call:
lm(formula = log(num_deaths + 1) ~ covid19 + allothercondandcauses +
  age25_34 + obesity + south + circulatorydiseases + alzheimer +
  vascular + malignant + covid19 * age25_34 + allothercondandcauses *
  age25_34 + obesity * age25_34 + circulatorydiseases * age25_34 +
  alzheimer * age25_34 + vascular * age25_34 + malignant *
  age25_34 + covid19 * south + allothercondandcauses * south,
  data = covid19dataUSA)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7840 -0.2022 -0.0256  0.0479  3.8149 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.02559   0.04311  0.594  0.552844    
covid19      1.28285   0.17831  7.195 9.93e-13 ***  
allothercondandcauses 0.90920   0.18115  5.019 5.83e-07 ***  
age25_34     0.61668   0.06321  9.755 < 2e-16 ***  
obesity      0.17356   0.14090  1.232  0.218247    
south        0.17660   0.04416  3.999 6.68e-05 ***  
circulatorydiseases -0.07351   0.06078 -1.209 0.226704    
alzheimer    -0.08446   0.11289 -0.748 0.454486    
vascular     -0.08564   0.11387 -0.752 0.452136    
malignant    -0.08937   0.13171 -0.679 0.497557    
covid19:age25_34 0.85885   0.21754  3.948 8.25e-05 ***  
allothercondandcauses:age25_34 0.65613   0.21226  3.091 0.002032 **  
age25_34:obesity 0.69414   0.24421  2.842 0.004539 **  
age25_34:circulatorydiseases -0.45179   0.09327 -4.844 1.41e-06 ***  
age25_34:alzheimer -0.61668   0.16141 -3.821 0.000139 ***  
age25_34:vascular -0.61668   0.16277 -3.789 0.000158 ***  
age25_34:malignant -0.62043   0.19017 -3.263 0.001130 **  
covid19:south    0.63402   0.21736  2.917 0.003589 **  
allothercondandcauses:south 0.72525   0.21077  3.441 0.000596 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.75 on 1475 degrees of freedom
Multiple R-squared:  0.4283,    Adjusted R-squared:  0.4214 
F-statistic:  61.4 on 18 and 1475 DF,  p-value: < 2.2e-16

```

Here, after removing outliers, I got best interaction model with log transformation for number of covid19 deaths. Looking at the F-test, p-value was below alpha level (0.05) we rejected the null hypothesis that all the betas are equal to zero and accept the alternative that at least one of the betas is not equal to zero. All the interaction terms are significant, so even though some of the independent variables were insignificant we need to include it because it is children of the interaction term. The adjusted R-square value is 42.14 percent which means our model predict 42.14% variability in response variable.

In the final interaction model, people of the age group 25-34 with the obesity, people from south region having more covid19 diseases and all other condition and causes, are contributing to increase in the number of covid19 deaths when comparing with Midwest region, Respiratory failure condition group and age group 0-24.

## **Sub section author student id: 2020637**

### **Based on M5:**

In milestone1 I focused on selecting my final explanatory variables for my first order model plus evaluating the model with the following parameters Adjusted R –Squared, F- values and the individual tests. I also validated my model with N fold cross validation which was useful in evaluating the MSE.

I also implemented both backward and forward selection processes, which I compared with my first order model. During the process of building my model, I encountered multicollinearity.

I loaded the variance packages

```
## Load packages  
```{r}  
library(tidyverse)  
library(gtsummary)
```

```
library(tableone)
library(broom)
```
``
```

Loaded my two datasets into R Studio

```
## Load data sets
```{r}
COVID <- read.csv("C:/Users/rejalu1/OneDrive - Henry Ford Health
System/DSC423/DataSets/COVID.csv")

states <- read.csv("C:/Users/rejalu1/OneDrive - Henry Ford Health
System/DSC423/DataSets/states.csv")
```
``
```

I selected variables of interest

```
## Select variables of interest
```{r}
covid_select <- COVID %>%
  select(state = State
        ,condition_group = Condition.Group
        ,condition = Condition
        ,age_group = Age.Group
        ,num_deaths = Number.of.COVID.19.Deaths)
```
``
```

I performed data cleaning on my dataset and also a performed a join on the states dataset to look up the regions variable using the state variable from the covid\_select data set and state\_code from the states data set.

```
## data cleaning
```{r}
covid_select$condition <-
str_replace(str_replace_all(str_replace_all(str_replace_all(str_replace_all(
str_replace_all(str_replace_all(str_replace_all(str_replace_all(str_replace_all(
str_replace_all(str_replace_all(str_replace_all(str_replace_all(str_replace_all(
str_replace_all(str_replace_all(str_replace_all(str_replace_all(str_replace_all(
covid_select$condition, 'Influenza and pneumonia', 'Influenza'), 'Chronic lower respiratory
diseases', 'ChronicLR'), 'Adult respiratory distress
syndrome', 'AdultRespDistSyn'), 'Respiratory
failure', 'Respiratoryfail'), 'Respiratory arrest', 'RespiratoryArrest'), 'Other
diseases of the respiratory system', 'OtherRespSystem'), 'Hypertensive
diseases', 'Hypertensive'), 'Ischemic heart disease', 'Ischemicheart'), 'Cardiac
arrest', 'CardiacArrest'), 'Cardiac arrhythmia', 'CardiacArrhythmia'), 'Heart
failure', 'HeartFailure'), 'Cerebrovascular diseases', 'Cerebrovascular'), 'Other
diseases of the circulatory system', 'OtherCirculatorySystem'), 'Malignant
neoplasms', 'Malignant'), 'Alzheimer disease', 'Alzheimer'), 'Vascular and
unspecified dementia', 'Vascular'), 'Intentional and unintentional injury,
poisoning, and other adverse events', 'injuryPoisoning'), 'All other conditions
and causes (residual)', 'OtherConditions&Causes')

covid_clean <- covid_select %>%
filter(state != 'US') %>%
filter(age_group != 'Not stated') %>%
filter(age_group != 'All ages') %>%
filter(age_group %in% c('35-44', '45-54')) %>%
inner_join(states, by=c('state'='state_code')) %>%
```

```

transmute(condition_group = as.factor(condition_group) ## using transmute
to remove any ## un wanted variables

  , condition = as.factor(condition)
  , age_group = as.factor(age_group)
  , region = as.factor(region)
  , num_deaths = num_deaths) %>%
  mutate(condition_group = relevel(condition_group, ref = 'Respiratory
diseases')
  , condition = relevel(condition, ref = 'Respiratoryfail')
  , age_group = relevel(age_group, ref = '35-44')
  , region = relevel(region, ref = 'Midwest')) %>%
na.omit()
```

```

From the same code snippet above using the relevel() function, I did specify different factor

```

## features selection using all stepwise selection
## building the full model with all the explanatory variables
## Adjusted R-squared:  0.234

```{r}
fullmodel <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West, data = covidcleansed)
```

```

levels as references on condition group, condition, age\_group and region independent variables. This ensures the ordering of the levels of a factor so that the level specified by ref are first and the others are moved down. Used stepwise selection to select my expected explanatory variables; Initially I build a model with all the explanatory variables from my cleaned data set as indicated below:

Used the summary() function to inspect the results of the full model fitting, below were my results:

```

Call:
lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
    alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
    chroniclr + covid19 + diabetes + heartfailure + hypertensive +
    influenza + injuryPoisoning + ischemicheart + malignant +
    obesity + othercirculatorysystem + otherrespsystem + renalfailure +
    conditionrespiratoryarrest + sepsis + vascular + Northeast +
    South + West, data = covidcleansed)

Residuals:
    Min      Q1 Median      Q3     Max 
-147.7  -22.9   -5.8    8.4 1026.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  31.57     9.40   3.36  0.00080 ***  
`age45-54`    27.47     3.85   7.14  1.6e-12 ***  
adultrespdistSyn -22.32    12.23  -1.83  0.06818 .    
allothcondsandcauses  21.04    11.41   1.84  0.06535 .    
alzheimer     -56.10    10.92  -5.14  3.2e-07 ***  
cardiacArrest  -27.85    13.15  -2.12  0.03438 *    
cardiacarrhythmia -51.02    12.79  -3.99  7.0e-05 ***  
cerebrovascula -49.77    13.52  -3.68  0.00024 ***  
chroniclr      -48.24    12.85  -3.75  0.00018 ***  
covid19        83.82     11.37   7.37  3.0e-13 ***  
diabetes       -21.64    12.35  -1.75  0.07990 .    
heartfailure   -49.41    13.52  -3.66  0.00027 ***  
hypertensive    -25.37    12.71  -2.00  0.04613 *    
influenza      21.20     11.89   1.78  0.07475 .    
injuryPoisoning -49.60    12.58  -3.94  8.5e-05 ***  
ischemicheart   -45.56    13.16  -3.46  0.00055 ***  
malignant       -50.22    13.35  -3.76  0.00018 ***  
obesity         -29.45    12.64  -2.33  0.01996 *    
othercirculatorysystem -44.73    13.32  -3.36  0.00080 ***  
otherrespsystem -49.51     13.52  -3.66  0.00026 ***  
renalfailure    -40.64    12.59  -3.23  0.00128 **  
conditionrespiratoryarrest -51.85    13.18  -3.93  8.8e-05 ***  
sepsis          -33.41    12.58  -2.66  0.00800 **  
vascular        -54.43    11.26  -4.83  1.5e-06 ***  
Northeast       14.85     6.01    2.47  0.01358 *    
South           16.16     5.36    3.02  0.00261 **  
West            12.26     5.60    2.19  0.02888 *  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.3 on 1337 degrees of freedom
Multiple R-squared:  0.249, Adjusted R-squared:  0.234 
F-statistic: 17.1 on 26 and 1337 DF,  p-value: <2e-16

```

So looking at the F-test, the P-value is less than 2e-16. The P-value is the probability that

given the null hypothesis, that all the Betas associated with the independent variables are equal to zero. We would observe the data as extreme as it is. Since the P-value are very small, so we rejected the null hypothesis and accepted the alternative, that at least one of the Betas is not equal to 0. We don't know which Beta or they are all not equal to zero. It is not what the F-TEST tells us. This is a test of the model itself, which tells me that something in my model is working.

The adjusted R-squared is 0.234, meaning that 23.4 percent of the variability in the number of COVID-19 deaths is explained by the model though adjusted R- Squared looks horrible.

Looking at the P-values from the t-test; since age group is a dummy variable and we set our reference level to '35-44' . The P-value for age-group '45-54' is 1.6e-12. That is the p-value for the null hypothesis that beta1 equals zero. Because the P-Value is low, we rejected that null hypothesis and accepted the alternative that Beta-1 is not equal to zero and then use the estimation of  $31.57 + 27.471$ , which is equivalent to 59.04. The estimation of the age group '35-44' will be 31.57, which is the y intercept.

Also remember that we defined region as a dummy variable, the p-values from the t-tests all look good. Since the P-values are low, we are going to reject the null hypothesis and accept the alternative that the betas for Midwest, Northeast, West and South are not equal to zero then use their estimations of 31.57 for Midwest,  $14.85 + 31.57 = 46.42$  for Northeast,  $16.16 + 31.57 = 47.73$  for South,  $12.26 + 31.57 = 43.83$  for West.

The same thing that applies to condition with the exclusion of adult respiratory distress syndrome, all other conditions and causes, diabetes, and Intentional, unintentional injury and poisoning have low P-values so, we rejected the null hypothesis and accepted the alternative that the betas associated with them are not equal to zero and then used their estimations. If we were to exclude these conditions, adult respiratory distress syndrome, all other conditions and causes, diabetes, and Intentional, unintentional injury and poisoning, whose P-values are high from the model. Their exclusion could impact.

the adjusted R-Squared negatively by reducing it and it could also affect the t-tests of other conditions which are considered significant. All the coefficients of the condition group were showing up as NA implying that there was multicollinearity in my data set, so I decided to drop it from my model.

Still using stepwise selection, I removed the following variables, diabetes and Adult Respiratory distress syndrome, which had the worst P-values one at a time until I got a nice clean first order model.

```

Call:
lm(formula = num_deaths ~ `age45-54` + allothcondsandcauses +
    alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
    chroniclr + covid19 + heartfailure + hypertensive + influenza +
    injuryPoisoning + ischemicheart + malignant + obesity + othercirculatorysystem +
    otherrespsystem + renalfailure + conditionrespiratoryarrest +
    sepsis + vascular + Northeast + South + West, data = covidcleansed)

Residuals:
    Min      1Q  Median      3Q     Max 
-147.4   -23.0   -5.4    8.2 1026.0 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.77     6.70    2.65  0.00812 **  
`age45-54`    27.35     3.85    7.10  2.0e-12 ***  
allothcondsandcauses  34.93     9.28    3.76  0.00018 ***  
alzheimer     -42.21     8.68   -4.86  1.3e-06 ***  
cardiacArrest   -13.95    11.36   -1.23  0.21969    
cardiacarrhythmia -37.14    10.94   -3.39  0.00071 ***  
cerebrovascula  -35.89    11.79   -3.04  0.00237 **  
chroniclr       -34.34    11.01   -3.12  0.00185 **  
covid19          97.70     9.25   10.57 < 2e-16 ***  
heartfailure    -35.54    11.79   -3.01  0.00263 **  
hypertensive     -11.45    10.84   -1.06  0.29092    
influenza        35.09     9.88    3.55  0.00039 ***  
injuryPoisoning -35.71    10.69   -3.34  0.00087 ***  
ischemicheart    -31.68    11.37   -2.79  0.00543 **  
malignant         -36.31    11.58   -3.13  0.00176 **  
obesity           -15.54    10.76   -1.44  0.14892    
othercirculatorysystem -30.84    11.55   -2.67  0.00770 **  
otherrespsystem   -35.62    11.78   -3.02  0.00255 **  
renalfailure      -26.73    10.70   -2.50  0.01261 *  
conditionrespiratoryarrest -37.95    11.39   -3.33  0.00089 ***  
sepsis            -19.50    10.69   -1.82  0.06834 .  
vascular          -40.55     9.11   -4.45  9.2e-06 ***  
Northeast          14.56     6.01    2.42  0.01560 *  
South              16.20     5.36    3.02  0.00256 **  
West               12.30     5.61    2.19  0.02840 *  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.4 on 1339 degrees of freedom
Multiple R-squared:  0.247, Adjusted R-squared:  0.233 
F-statistic: 18.3 on 24 and 1339 DF,  p-value: <2e-16

```

My F-test-remained the same, Adjusted R-squared remained the same. Since my adjusted R-squared was not increasing, I decided to stop here for my first order model. When I used both backward selection to select the expected explanatory

```
## Model selection

## building a full model with everything using the backward selection
```{r}
library(MASS
)
#model_full <- lm(num_deaths ~ age_group + condition + condition_group
+ region, data = covid_clean)

model_full <- lm(num_deaths ~ ., data =
covidcleansed) step <- stepAIC(model_full, direction
= "backward") step$anova ## display results
summary(step)
```

```

I end up getting the same final model as indicated below after running the StepAIC():

```

Start:  AIC=11629
num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West

              Df Sum of Sq     RSS    AIC
<none>                      6612665 11629
- diabetes                   1    15192 6627858 11630
- influenza                  1    15730 6628396 11631
- adultrespdistSyn          1    16478 6629144 11631
- allothcondsandcauses      1    16824 6629489 11631
- hypertensive                1    19706 6632371 11631
- cardiacArrest              1    22183 6634848 11632
- West                        1    23668 6636333 11632
- obesity                     1    26848 6639514 11633
- Northeast                   1    30210 6642876 11634
- sepsis                      1    34888 6647553 11635
- South                       1    44999 6657665 11637
- renalfailure                1    51547 6664212 11638
- othercirculatorysystem      1    55804 6668470 11639
- ischemicheart               1    59294 6671960 11640
- heartfailure                 1    66076 6678741 11641
- otherrespsystem              1    66370 6679035 11641
- cerebrovascula              1    67068 6679733 11641
- chroniclr                    1    69698 6682363 11642
- malignant                    1    70021 6682687 11642
- conditionrespiratoryarrest   1    76533 6689199 11643
- injuryPoisoning              1    76887 6689552 11643
- cardiacarrhythmia            1    78757 6691422 11643
- vascular                     1    115542 6728207 11651
- alzheimer                    1    130494 6743159 11654
- `age45-54`                   1    251833 6864498 11678
- covid19                      1    268642 6881307 11682

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +

```

```

South + West

Final Model:
num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West

Step Df Deviance Resid. Df Resid. Dev AIC
1 1337 6612665 11629

Call:
lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West, data = covidcleansed)

Residuals:
    Min      1Q Median      3Q     Max
-147.7   -22.9   -5.8    8.4 1026.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.57      9.40   3.36  0.00080 ***
`age45-54`  27.47      3.85   7.14  1.6e-12 ***
adultrespdistSyn -22.32     12.23  -1.83  0.06818 .
allothcondsandcauses 21.04     11.41   1.84  0.06535 .
alzheimer -56.10     10.92  -5.14  3.2e-07 ***
cardiacArrest -27.85     13.15  -2.12  0.03438 *
cardiacarrhythmia -51.02     12.79  -3.99  7.0e-05 ***
cerebrovascula -49.77     13.52  -3.68  0.00024 ***
chroniclr -48.24     12.85  -3.75  0.00018 ***
covid19  83.82     11.37   7.37  3.0e-13 ***
diabetes -21.64     12.35  -1.75  0.07990 .
heartfailure -49.41     13.52  -3.66  0.00027 ***
hypertensive -25.37     12.71  -2.00  0.04613 *
influenza  21.20     11.89   1.78  0.07475 .
injuryPoisoning -49.60     12.58  -3.94  8.5e-05 ***
ischemicheart -45.56     13.16  -3.46  0.00055 ***
malignant -50.22     13.35  -3.76  0.00018 ***
obesity -29.45     12.64  -2.33  0.01996 *
othercirculatorysystem -44.73     13.32  -3.36  0.00080 ***
otherrespsystem -49.51     13.52  -3.66  0.00026 ***

```

using backward selection, my P-values from the F-test, Adjusted R-Squared and the P-values from the t- tests are still the same in as in my initial full model initially built on page

5.

I also used forward selection as defined below, built an empty model and ran the stepAIC() function, which Ended up with same final model as below:

```
## building a full model with everything using foward selection
```{r}
model_full <- lm(num_deaths ~ ., data = covidcleansed)
model_empty <- lm(num_deaths ~ 1, data = covidcleansed)
summary(model_empty)
## pass the empty model to stepAIC()
## consider every variable all the way down to no variable
## that is why we define the scope

step <- stepAIC(model_empty, direction = "forward", scope =
list(upper=model_full, lower=model_empty))
summary(step)

Call:
lm(formula = num_deaths ~ covid19 + `age45-54` + allothcondsandcauses +
influenza + alzheimer + South + vascular + diabetes + adultrespdistSyn +
Northeast + West + hypertensive, data = covidcleansed)

Residuals:
    Min      1Q  Median      3Q     Max 
-147.4   -23.4   -5.4    9.6 1024.3 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.59      4.94   -1.94  0.05228 .  
covid19       123.67     8.20   15.08 < 2e-16 *** 
`age45-54`    28.52      3.86    7.39  2.5e-13 *** 
allothcondsandcauses 60.92     8.24    7.39  2.5e-13 *** 
influenza     60.91     8.92    6.83  1.3e-11 ***
```

```

alzheimer          -16.14      7.54   -2.14  0.03244 *
South              18.12      5.37   3.38  0.00076 ***
vascular           -14.40      8.02   -1.79  0.07303 .
diabetes            18.08      9.52   1.90  0.05784 .
adultrespdistSyn  17.52      9.36   1.87  0.06148 .
Northeast           14.80      6.04   2.45  0.01442 *
West                12.42      5.63   2.21  0.02750 *
hypertensive        ... ...
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70.9 on 1351 degrees of freedom
Multiple R-squared:  0.23,  Adjusted R-squared:  0.223
F-statistic: 33.6 on 12 and 1351 DF,  p-value: <2e-16

```

Using forward selection, my P-values from the F-test remained the same, Adjusted R-Squared dropped by 0.11, which was not big of a change . Using the forward selection model most of the variables like renalFailure, malignant, ischemiheart e.tc were found to be insignificant but I wouldn't consider this model as accurate as it may look like, however, it can still be used.

So, with both all stepwise selection, backward and forward selections, I end up with same first order model that had the same F-test P values, Adjusted R-Squared and P values from the t-tests.

There was multicollinearity in the dataset being caused by the overlap between conditions and condition group. I removed condition group because it was one of the variables causing multicollinearity in the dataset. When I looked at the Condition groups and conditions in the COVID data set, they were some data hygiene issues, which needed to be corrected from the source system where this data was collected from. For validation purposes I took the covidcleansed data set and split it into an eighty-twenty split that is between the train and test data sets as below and took the initial full model, which I named model.covid19\_deathsA with all the independent variables, and also took and created another model with all the independent variables but pruned diabetes off and built these models using the train data sets and evaluated it using the test data set.

```

## divide your data set into train and test data sets
```{r}
set.seed(123)
partition <- sample(2,nrow(covidcleansed),replace = TRUE, prob= c(0.80,0.20))
train <- covidcleansed[partition==1, ]
test <- covidcleansed[partition==2, ]

library(MASS)
##data = covidcleanse
model.covid19_deathsA <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West, data=train)
summary(model.covid19_deathsA)
prediction <- predict(model.covid19_deathsA,test)
actual =test$num_deaths

```

```

cor(prediction, actual)

model.covid19_deathsB <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West, data=train)
summary(model.covid19_deathsB)
prediction <- predict(model.covid19_deathsB,test)
actual =test$num_deaths

```

```

cor(prediction, actual)
#plot(prediction.actual)
```

model.covid19\_deathsA gave me an adjusted R-squared of 0.221 For model A, the correlation between prediction and actual is 0.526

model.covid19\_deathsB gave me an adjusted R-squared of 0.219 For model B, the correlation between prediction and actual is 0.53.

I also used the N fold cross validation two validate my two models.

```
outA <- cv.lm(data = test
                , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West)
                , plotit = "Observed", m=5)

outB <- cv.lm(data = test
                , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West)
                , plotit = "Observed", m=5)
```

Model.covid19\_deathsA had an average mean square error of 5314 while model.covid19\_deathsB had an average mean square error of 5311.

In milestone2 I worked on transforming my variables, analyzing the need for second order terms, interaction terms, and also checking for multicollinearity. I performed residual analysis in order to stabilize my final model.

```

## Second Order and Interaction terms

`~`{r}
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interaction <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  adultrespdistSyn*`age45-54` + allothcondsandcauses*`age45-54` +
  alzheimer*`age45-54` + cardiacArrest*`age45-54` +
  cardiacarrhythmia*`age45-54` + cerebrovascula*`age45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` +
  heartfailure*`age45-54` + hypertensive*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-
54` + malignant*`age45-54` +
  obesity*`age45-54` + othercirculatorysystem*`age45-54` +
  otherrespsystem*`age45-54` + renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + sepsis*`age45-54` +
  vascular*`age45-54` + Northeast*`age45-54` +
  South*`age45-54` + West*`age45-54` , data = covidcleansed)

```

```
summary(model.interaction)
```

I created a second order term as defined above, built and ran my model, and ended up getting NA Coefficients for second order term as indicated below in the summary statistics results. Furthermore, when I added interaction terms combining conditions and age group and my plan, I used stepwise selection to select the most significant interaction terms. So, running the summary() function, this is what I got:

```

Call:
lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
    alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
    chroniclr + covid19 + diabetes + heartfailure + hypertensive +
    influenza + injuryPoisoning + ischemicheart + malignant +
    obesity + othercirculatorysystem + otherrespsystem + renalfailure +
    conditionrespiratoryarrest + sepsis + vascular + Northeast +
    South + West + + `age45-54SQ` + covid19SQ + adultrespdistSyn *
    `age45-54` + allothcondsandcauses * `age45-54` + alzheimer *
    `age45-54` + cardiacArrest * `age45-54` + cardiacarrhythmia *
    `age45-54` + cerebrovascula * `age45-54` + chroniclr * `age45-54` +
    covid19 * `age45-54` + diabetes * `age45-54` + heartfailure *
    `age45-54` + hypertensive * `age45-54` + influenza * `age45-54` +
    injuryPoisoning * `age45-54` + ischemicheart * `age45-54` +
    malignant * `age45-54` + obesity * `age45-54` + othercirculatorysystem *
    `age45-54` + otherrespsystem * `age45-54` + renalfailure *
    `age45-54` + conditionrespiratoryarrest * `age45-54` + sepsis *
    `age45-54` + vascular * `age45-54` + Northeast * `age45-54` +
    South * `age45-54` + West * `age45-54`, data = covidcleansed)

Residuals:
    Min      Q1 Median      Q3     Max 
-195.8   -17.3   -4.6    4.4   981.6 

Coefficients: (2 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.90     13.27   1.80   0.0720 .  

```

|                            |        |       |       |        |    |
|----------------------------|--------|-------|-------|--------|----|
| age 45 - 54                | 39.68  | 18.08 | 2.20  | 0.0283 | •  |
| adultrespdist syn          | -12.13 | 17.58 | -0.69 | 0.4903 |    |
| allothcondsandcauses       | 11.23  | 16.16 | 0.69  | 0.4873 |    |
| alzheimer                  | -29.69 | 15.44 | -1.92 | 0.0547 |    |
| cardiacArrest              | -13.64 | 19.32 | -0.71 | 0.4801 |    |
| cardiacarrhythmia          | -28.23 | 17.45 | -1.62 | 0.1061 |    |
| cerebrovascula             | -28.52 | 18.13 | -1.57 | 0.1158 |    |
| chroniclr                  | -26.30 | 18.13 | -1.45 | 0.1472 |    |
| covid19                    | 50.09  | 16.26 | 3.08  | 0.0021 | •• |
| diabetes                   | -6.61  | 19.03 | -0.35 | 0.7284 |    |
| heartfailure               | -28.27 | 17.95 | -1.57 | 0.1155 |    |
| hypertensive               | -13.15 | 18.77 | -0.70 | 0.4837 |    |
| influenza                  | 14.14  | 17.14 | 0.83  | 0.4095 |    |
| injuryPoisoning            | -25.71 | 17.75 | -1.45 | 0.1477 |    |
| ischemicheart              | -27.35 | 17.78 | -1.54 | 0.1243 |    |
| malignant                  | -27.57 | 18.36 | -1.50 | 0.1335 |    |
| obesity                    | -5.87  | 18.54 | -0.32 | 0.7517 |    |
| othercirculatorysystem     | -24.62 | 18.76 | -1.31 | 0.1896 |    |
| otherrespsystem            | -26.85 | 18.55 | -1.45 | 0.1479 |    |
| renalfailure               | -18.47 | 19.90 | -0.93 | 0.3536 |    |
| conditionrespiratoryarrest | -27.05 | 18.41 | -1.47 | 0.1419 |    |

|           |        |       |       |        |
|-----------|--------|-------|-------|--------|
| sepsis    | -19.38 | 18.13 | -1.07 | 0.2853 |
| vascular  | -29.62 | 15.57 | -1.90 | 0.0573 |
| Northeast | 6.28   | 8.58  | 0.73  | 0.4644 |
| South     | 10.33  | 7.63  | 1.35  | 0.1761 |
| West      | 4.86   | 7.65  | 0.64  | 0.5250 |

|                                        |        |          |        |           |
|----------------------------------------|--------|----------|--------|-----------|
| `age45-54SQ`                           | NA     | NA       | NA     | NA        |
| covid19SQ                              | NA     | NA       | NA     | NA        |
| `age45-54` :adultrespdistSyn           | -19.18 | 24.07    | -0.80  | 0.4257    |
| `age45-54` :allothcondsandcauses       | 21.31  | 22.43    | 0.95   | 0.3422    |
| `age45-54` :alzheimer                  | -52.28 | 21.48    | -2.43  | 0.0151 *  |
| `age45-54` :cardiacArrest              | -27.71 | 26.00    | -1.07  | 0.2866    |
| `age45-54` :cardiacarrhythmia          | -48.54 | 25.27    | -1.92  | 0.0550 .  |
| `age45-54` :cerebrovascula             | -46.55 | 26.85    | -1.73  | 0.0832 .  |
| `age45-54` :chroniclr                  | -43.32 | 25.27    | -1.71  | 0.0867 .  |
| `age45-54` :covid19                    | 65.33  | 22.38    | 2.92   | 0.0036 ** |
| `age45-54` :diabetes                   | -27.65 | 24.77    | -1.12  | 0.2644    |
| `age45-54` :heartfailure               | -47.26 | 27.02    | -1.75  | 0.0805 .  |
| `age45-54` :hypertensive               | -24.27 | 25.15    | -0.97  | 0.3347    |
| `age45-54` :influenza                  | 13.01  | 23.41    | 0.56   | 0.5786    |
| `age45-54` :injuryPoisoning            | -46.62 | 24.74    | -1.88  | 0.0598 .  |
| `age45-54` :ischemicheart              | -37.27 | 26.09    | -1.43  | 0.1534    |
| `age45-54` :malignant                  | -47.74 | 26.33    | -1.81  | 0.0700 .  |
| `age45-54` :obesity                    | -44.31 | 24.97    | -1.77  | 0.0762 .  |
| `age45-54` :othercirculatorysystem     | -38.67 | 26.18    | -1.48  | 0.1398    |
| `age45-54` :otherrespsystem            | -47.11 | 26.65    | -1.77  | 0.0773 .  |
| `age45-54` :renalfailure               | -39.02 | 25.52    | -1.53  | 0.1265    |
| `age45-54` :conditionrespiratoryarrest | -50.73 | 25.95    | -1.96  | 0.0508 .  |
| `age45-54` :sepsis                     | -26.65 | 24.77    | -1.08  | 0.2821    |
| `age45-54` :vascular                   | -51.81 | 22.22    | -2.33  | 0.0199 *  |
| `age45-54` :Northeast                  | 20.49  | 11.87    | 1.73   | 0.0847 .  |
| `age45-54` :South                      | 14.04  | 10.58    | 1.33   | 0.1848    |
| `age45-54` :West                       | 17.68  | 11.06    | 1.60   | 0.1101    |
| ---                                    |        |          |        |           |
| Signif. codes:                         | 0 ***  | 0.001 ** | 0.01 * | 0.05 .    |
|                                        | 0.1    | 1        |        |           |

```
Residual standard error: 69.1 on 1312 degrees of freedom
Multiple R-squared:  0.289, Adjusted R-squared:  0.262
F-statistic: 10.5 on 51 and 1312 DF,  p-value: <2e-16
```

The NA coefficients showed a sign of multicollinearity in my data set being introduced by the addition of the second order terms so, I ended up excluding both age45-54SQ and

covid19SQ.

Used N fold cross validation to train and tested 8 different models using my interaction terms to get the best model.

| Model name         | Adjusted R-Squared | Average Mean Square Error |
|--------------------|--------------------|---------------------------|
| Model.interactionA | 0.252              | 5517                      |
| Model.interactionB | 0.253              | 5442                      |
| Model.interactionC | 0.253              | 5427                      |
| Model.interactionD | 0.254              | 5428                      |
| Model.interactionE | 0.254              | 5375                      |
| Model.interactionF | 0.255              | 5370                      |
| Model.interactionG | 0.255              | 5375                      |
| Model.interactionH | 0.26               | 5355                      |

So, building and running model.interactionH on my covidcleansed dataset

```
## Final model with interaction terms
## Pruned`age45-54`:chroniclr , `age45-54`:obesity
##, `age45-54`:heartfailure, `age45-54`:cerebrovascula
##, ##`age45-54`:otherrespsystem, `age45-54`:malignant , `age45-
54`:injuryPoisoning, `age45-54`:cardiacarrhythmia, ##`age45-
54`:conditionrespiratoryarrest, `age45-54`:vascular
##, age45-54`:alzheimer , `age45-54`:renalfailure, `age45-
54`:othercirculatorysystem
``{r}
covidcleansed$`age45-54SQ`<- covidcleansed$`age45-54`^2
model.interactionH <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ## |
  allothcondsandcauses*`age45-54` + covid19*`age45-54` +
  influenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` +
  West*`age45-54`, data = covidcleansed)
summary(model.interactionH)
``
```

Running the Summary function on this final model

```

Call:
lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses * `age45-54` + covid19 *
`age45-54` + influenza * `age45-54` + Northeast * `age45-54` +
South * `age45-54` + West * `age45-54`, data = covidcleansed)

```

Residuals:

| Min    | 1Q    | Median | 3Q  | Max   |
|--------|-------|--------|-----|-------|
| -195.1 | -20.3 | -4.3   | 6.9 | 980.9 |

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(> t )    |
|----------------------|----------|------------|---------|-------------|
| (Intercept)          | 45.499   | 10.028     | 4.54    | 6.2e-06 *** |
| `age45-54`           | 0.231    | 8.411      | 0.03    | 0.97812     |
| adultrespdistSyn     | -22.422  | 11.974     | -1.87   | 0.06134 .   |
| allothcondsandcauses | -9.154   | 13.698     | -0.67   | 0.50410     |
| alzheimer            | -57.026  | 10.698     | -5.33   | 1.1e-07 *** |
| cardiacArrest        | -28.234  | 12.889     | -2.19   | 0.02866 *   |
| cardiacarrhythmia    | -53.182  | 12.531     | -4.24   | 2.3e-05 *** |
| cerebrovascula       | -52.340  | 13.247     | -3.95   | 8.2e-05 *** |
| chroniclr            | -49.287  | 12.588     | -3.92   | 9.5e-05 *** |
| covid19              | 29.636   | 13.811     | 2.15    | 0.03207 *   |

|                                  |         |        |       |         |     |
|----------------------------------|---------|--------|-------|---------|-----|
| diabetes                         | -20.360 | 12.092 | -1.68 | 0.09247 | .   |
| heartfailure                     | -52.240 | 13.251 | -3.94 | 8.5e-05 | *** |
| hypertensive                     | -25.599 | 12.451 | -2.06 | 0.03997 | *   |
| influenza                        | -6.130  | 14.846 | -0.41 | 0.67973 |     |
| injuryPoisoning                  | -50.232 | 12.326 | -4.08 | 4.9e-05 | *** |
| ischemicheart                    | -47.497 | 12.899 | -3.68 | 0.00024 | *** |
| malignant                        | -52.355 | 13.087 | -4.00 | 6.7e-05 | *** |
| obesity                          | -29.751 | 12.381 | -2.40 | 0.01640 | *   |
| othercirculatorysystem           | -45.236 | 13.041 | -3.47 | 0.00054 | *** |
| otherrespsystem                  | -51.274 | 13.243 | -3.87 | 0.00011 | *** |
| renalfailure                     | -39.397 | 12.332 | -3.19 | 0.00143 | **  |
| conditionrespiratoryarrest       | -53.571 | 12.921 | -4.15 | 3.6e-05 | *** |
| sepsis                           | -33.596 | 12.318 | -2.73 | 0.00647 | **  |
| vascular                         | -56.005 | 11.032 | -5.08 | 4.4e-07 | *** |
| Northeast                        | 4.799   | 8.520  | 0.56  | 0.57337 |     |
| South                            | 8.604   | 7.564  | 1.14  | 0.25556 |     |
| West                             | 3.640   | 7.600  | 0.48  | 0.63203 |     |
| `age45-54` :allothcondsandcauses | 59.403  | 15.780 | 3.76  | 0.00017 | *** |
| `age45-54` :covid19              | 103.492 | 15.705 | 6.59  | 6.3e-11 | *** |
| `age45-54` :influenza            | 50.918  | 17.165 | 2.97  | 0.00307 | **  |
| `age45-54` :Northeast            | 21.467  | 11.802 | 1.82  | 0.06916 | .   |
| `age45-54` :South                | 16.661  | 10.479 | 1.59  | 0.11208 |     |
| `age45-54` :West                 | 18.362  | 10.981 | 1.67  | 0.09474 | .   |

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.9 on 1331 degrees of freedom
Multiple R-squared:  0.283, Adjusted R-squared:  0.266
F-statistic: 16.4 on 32 and 1331 DF, p-value: <2e-16
```

Interpreting the regression coefficients of the final model:

Looking at the F -test, the P-Value is than 2e-16. The P-Value is the probability that given the null hypothesis, that all the Betas associated with the independent variables are equal to zero. We would observe the data as extreme as it is. Since the P-value was very small, so we rejected the null hypothesis and accepted the alternative, that at least one of the Betas was not equal to 0. We don't know

which Beta or they are all not equal to zero. It is not what the F-TEST tells us. This is a test of the model itself, which tells me that something in my model was working. Adjusted R-squared is 0.266 meaning that 26.6 percent of the variability in the number of Covid-19 deaths is explained by the model.

Looking at the individual t-tests, the P-Values for Alzheimer, cardiacAre, cardiacArrest, cardiacarrhythmia, cerebrovascular, chroniclr, covid19, heartfailure, hypertensive, injuryPoisoning , ischemicheart, malignant, obesity, othercirculatorysystem, otherrespsystem, renalfailure , conditionrespiratoryarrest , sepsis , vascular , `age45-54`:allothcondsandcauses , age45-54`:covid19 , and `age45-54`:influenza, those are the p-values for the null hypothesis that the Betas associated with those variables are equal to zero. Because the corresponding P-Values are low, we are going to reject the null hypothesis and accept the alternative that the Betas associated with those variables are not equal to zero and then use their estimations. Remember we set dummy variables on agegroup with a reference level of 35-44 as and also, on conditions with reference level of Respiratoryfail so, the estimation of age35-44 will by the y- intercept of 45.499, age45-54 will be  $45.499 + 0.231$ , Respiratoryfail will have an estimation of 45.499, Alzheimer will have  $45.499 + (-57.026)$ , cardiacArrest will be  $45.499 + (-28.234)$ , the interaction term age45-54:allothercondsandcauses will have an estimate ion of  $45.499 + 59.403$ , age45-54:covid19 will have an estimation of  $45.499 + 103.492$ , age45-54:influenza will have an estimation of  $45.499 + 50.918$

The P-values for age45-54, adultrespdistSyn, allothcondsandcauses, diabetes, influenza, Northeast, South, West, `age45-54`:Northeast, `age45-54`:South and `age45-54`:West are greater than 0.05, which is my default level of significance implying we never accept the null hypothesis that those betas associated with those variables are equal to zero hence we are un able to reject the null hypothesis so, something is definitely wrong in my model, so I plan to diagnose whatever the issue in my model is by looking at residuals.

Some of the p-values for the corresponding interaction terms `age45-54`:allothcondsandcauses' and `age45-54`:influenza are significant so their child parameters like age45-54, allothcondsandcauses, and Influenza will have to be included in the model regardless of their high P – values because we always keep the children when we include the interaction terms.

Residual Analysis:

```
## Residual Analysis
```

```

## Verifying the assumption that the sum of the residuals should add up
to zero.

```{r}
model.interactionH <- lm(num_deaths ~ `age45-54` + adultrespdistSyn
+ allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula
  + chroniclr + covid19 + diabetes + heartfailure + hypertensive
  + influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure
  + conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` +
  influenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` +
  West*`age45-54`, data = covidcleansed)
sum(model.interactionH$residuals)
```

```

The above code snippet above used the `sum()` function with a list of residuals passed as an argument to add up all the residuals to zero:

That gives you:

```
[1] -1.81e-14
```

This is an equivalent to zero, so this verifies the assumption that the sum of residuals add to zero. The code below defines the histogram of residuals:

```

##Histogram of residuals

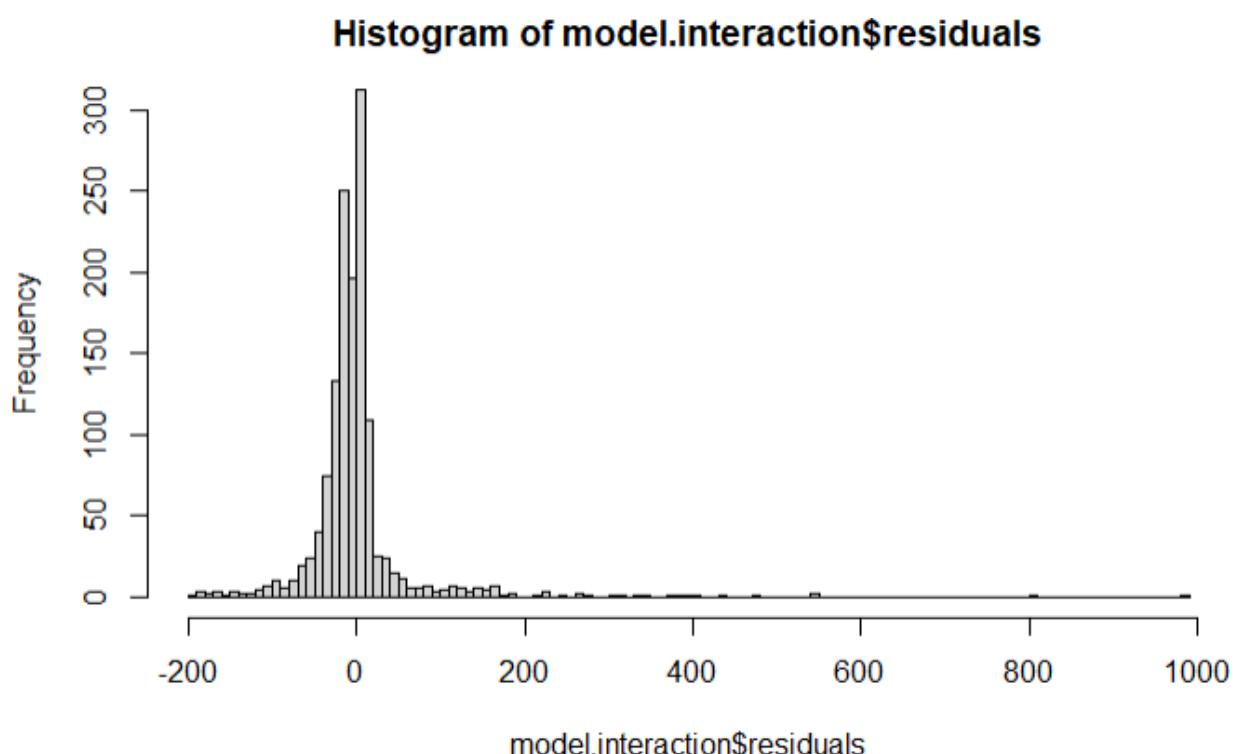
```{r}
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interaction <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
```

```

```

conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses` age45-54` + covid19` age45-54` +
+ influenza` age45-54` +
+ Northeast` age45-54` + South` age45-54` + West` age45-54` , data
= covidcleansed)
hist(model.interaction$residuals, breaks = 100)
```

```



This is an iconic normal curve with a tall peak. It also has little tails on the left and right hand

side. This verifies the assumption that the residuals are normally distributed and indeed they are normally distributed. There are a number of outliers that we see. We don't know how far there are in relation to the standard deviation. So, we go ahead and get an idea about the distribution of the curve by taking the mean of the residuals, computing the standard deviation of the residuals, z-score normalization of the residuals computed by subtracting the standard deviation of the residual, which is a single number, from the list of residuals which results into a new list and each element in the list will be divided by the standard deviation. Normalized z-score residuals will be used to draw a histogram. This is demonstrated below.

```
##idea about the distribution in the curve
```

```
```{r}
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interaction <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West +
##
allothcondsandcauses*`age45-54` + covid19*`age45-54` +
influenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` +
West*`age45-54`, data = covidcleansed)

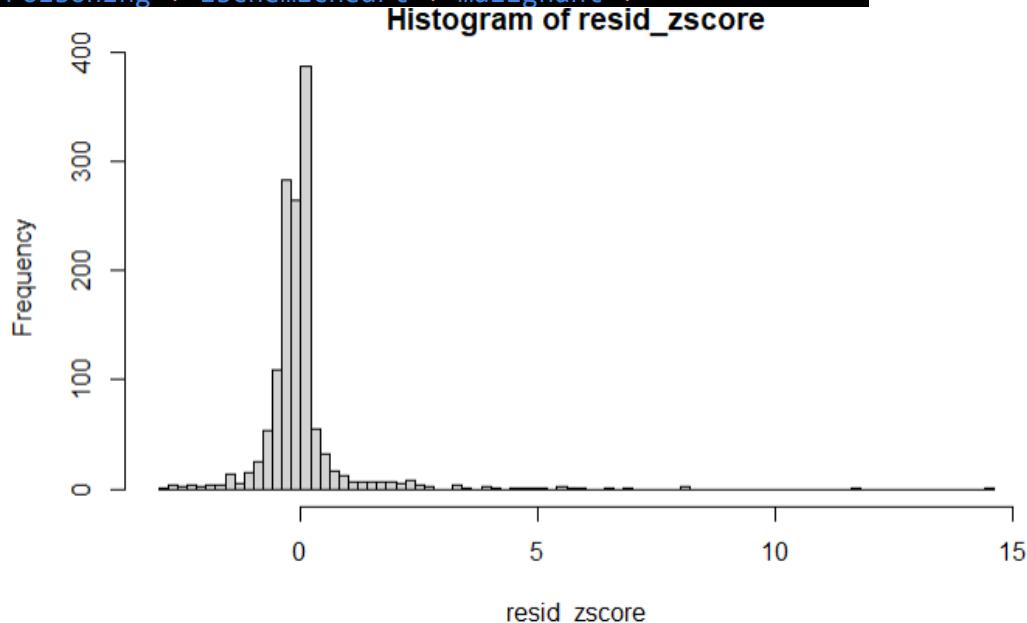
mean = mean(model.interaction$residuals) #

## This is the standard deviation of the residuals which I am to save as sd.
sd = sd(model.interaction$residuals)
resid_zscore = (model.interaction$residuals - mean)/sd
hist(resid_zscore, breaks = 100)
```
```

```
##Verifying the  
independence of residuals  
## using Durbin Watson test  
on the model.
```

```
```{r}
```

```
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-  
54`^2 model.interaction <- lm(num_deaths ~ `age45-54`  
+ adultrespdistSyn + allothcondsandcauses +  
alzheimer + cardiacArrest + cardiacarrhythmia +  
cerebrovascula + chroniclr + covid19 + diabetes +  
heartfailure + hypertensive + influenza +  
injuryPoisoning + ischemicheart + malignant +
```



As shown above, we expect that 95 percent of the residuals are within two standard deviations and there are outliers on the right something we need to be aware of. I used the Durbin Watson test to verify the assumption that the residuals should be independent:

```

durbinWatsonTest(model.interaction)
```
lag Autocorrelation D-W Statistic p-
value
1          0.42          1.16      0
Alternative hypothesis: rho != 0

```

Interpreting the results of the Durbin Watson test, a p-value of 0 casts doubt about the validity of the null hypothesis, which indicates auto correlation amongst the residuals.

Residual plots:

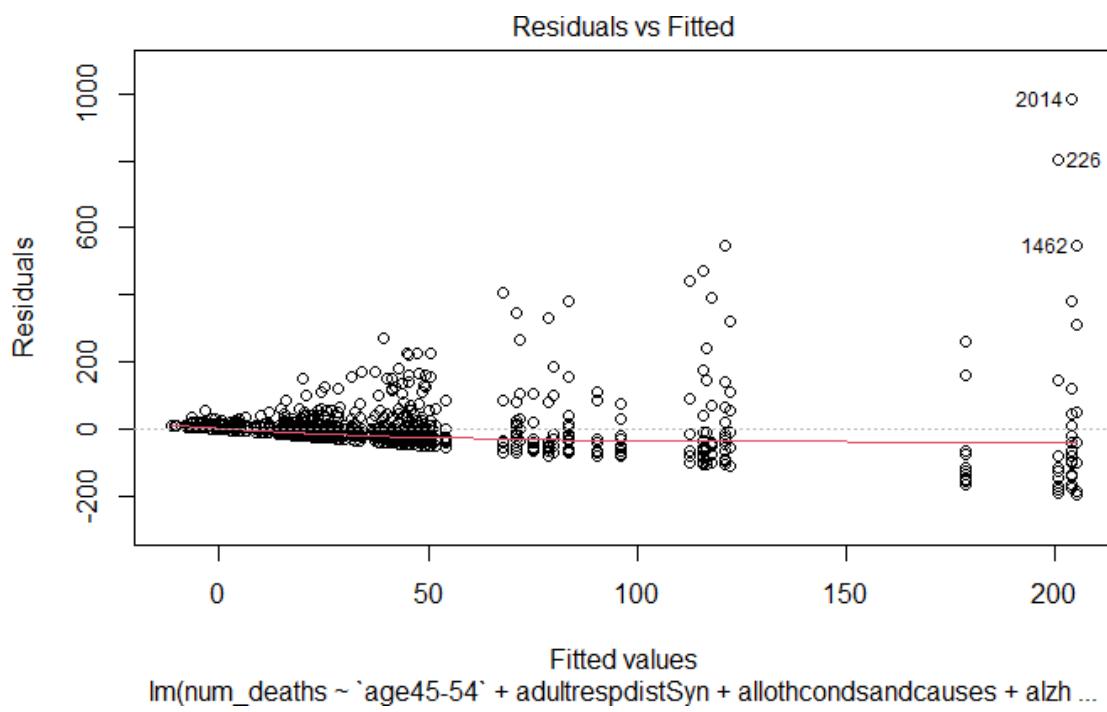
I used the plot() function to graph the different residuals plots to study the built model:

```

## Residual plots
```{r}
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` +
influenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` +
West*`age45-54`, data = covidcleansed)
prediction <- predict(model.interaction, covidcleansed)
plot(model.interactionH)
```

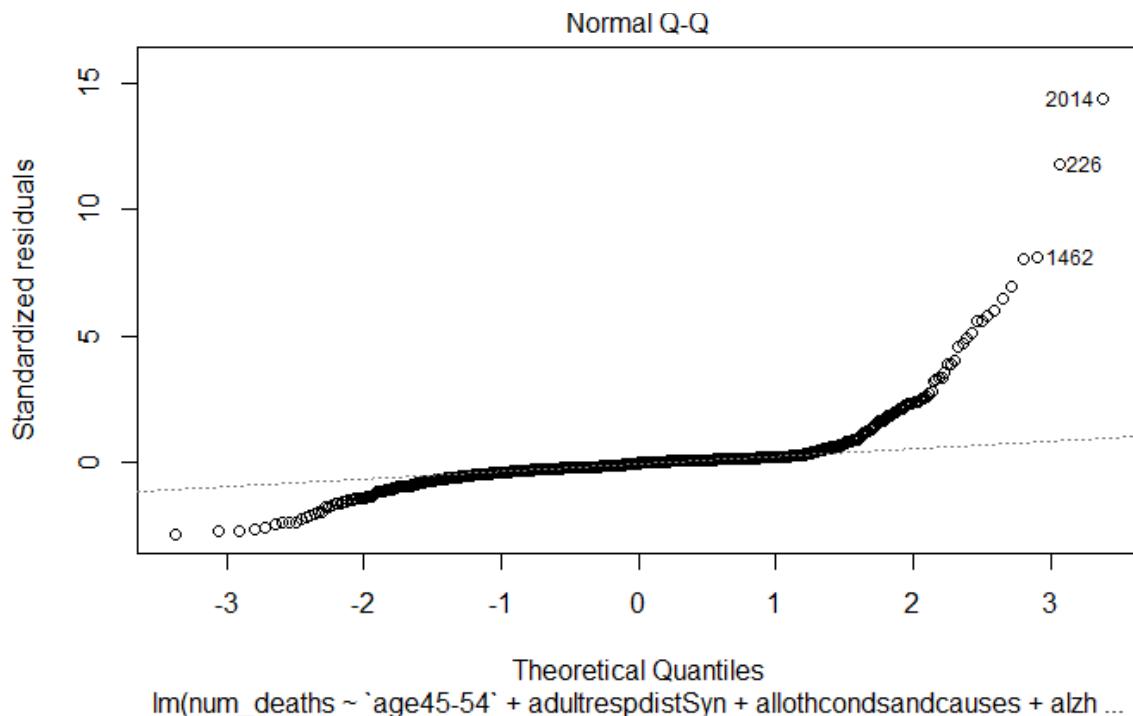
```

My first graph was the Residuals Vs Fitted

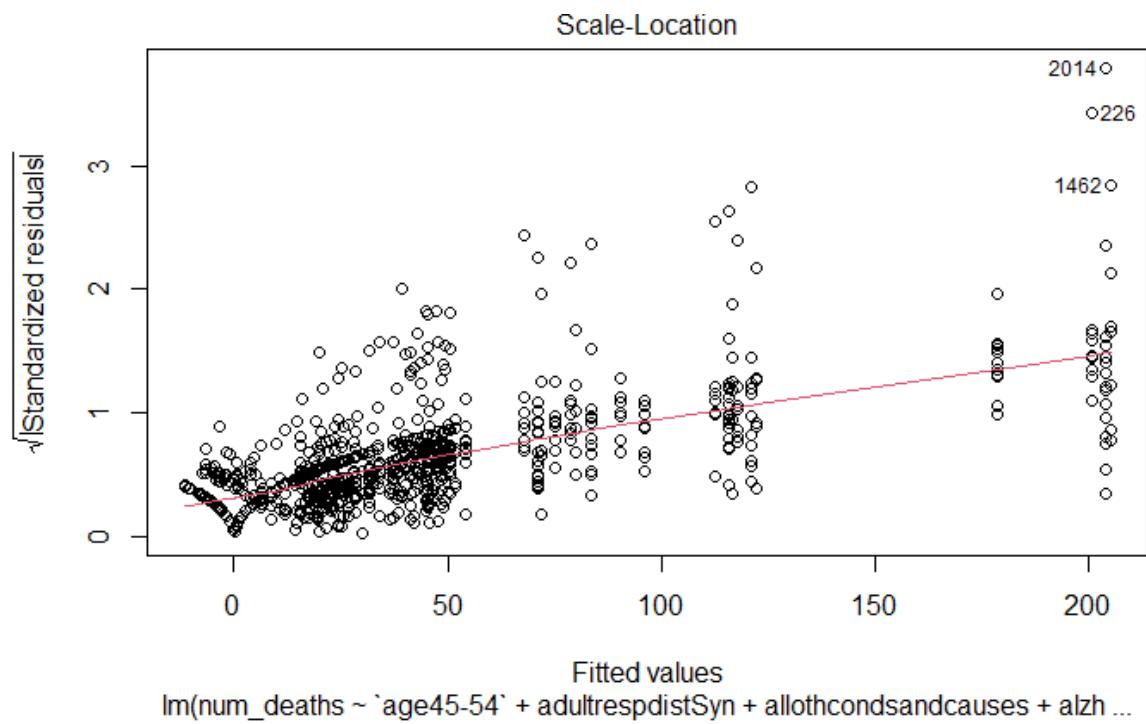


Looking at the above residual plot, there is a multiplicative pattern in the variance especially as you move from left to right so instead of predicting the number of death, we shall use the log of num\_death plus 1, that is to say,  $\log(\text{num\_death} + 1)$

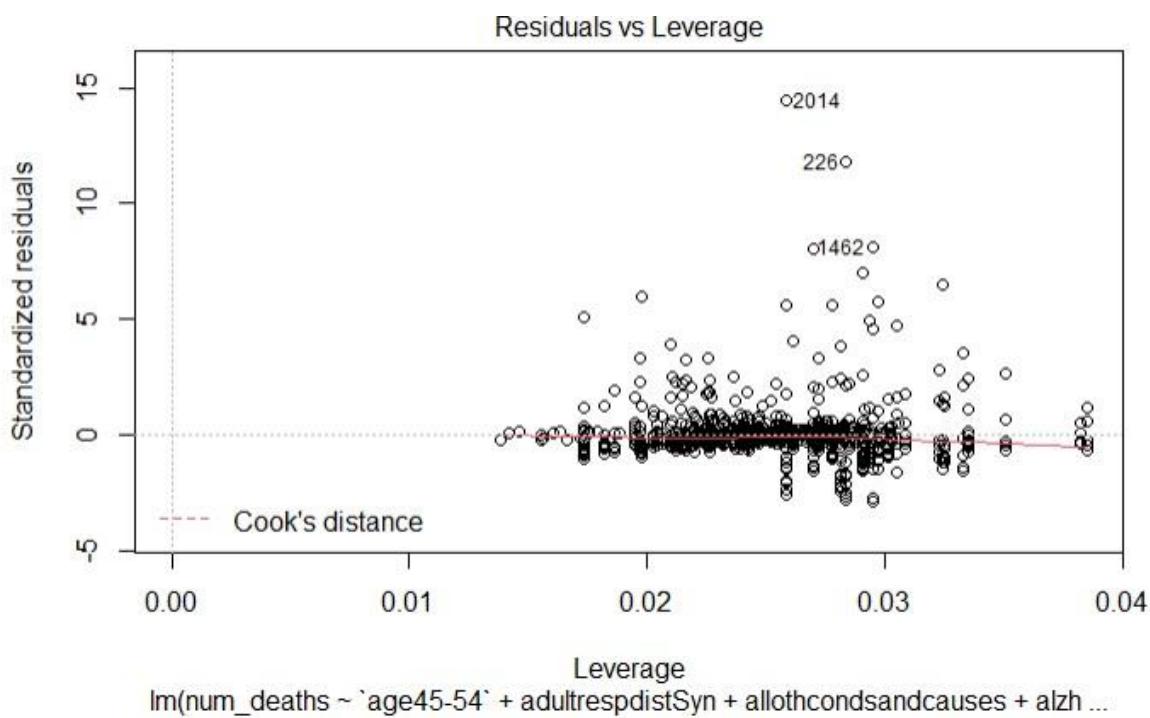
The Q-Q plot below looks shows that the residuals are quite normal.



The third one is fitted value against the square root of standardized residuals



The fourth residual plot shows us the leverage versus the standardized residuals. It labeled a few of the points of observations 2014, 226, 1462, which had lot of leverage on the regression line. So, we went into the data set, removed those observations so that the model could be stabilized.



Points 2014, 1462, 226 corresponded to observations 1203, 880, 153 respectively.

```
## Removing specific rows in r
```{r}
## Adding an index column to my data set
covidcleansed$generated_uid <- 1:nrow(covidcleansed)
covidcleansed <- covidcleansed[-c(1203,880,153),]
```

So, using the above code snippet, I removed those observations from the covidcleansed data set. To stabilize the model, instead of predicting of num\_deaths, I used the log(num\_deaths + 1) as indicated below:

```
## Stabilized model
```{r}
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2

model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
+ allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` +
  influenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` +
  West*`age45-54`, data = covidcleansed)
prediction <- predict(model.interaction,
covidcleansed) summary(model.interactionH)
##plot(model.interactionH)
```

```

Running the summary(model.interactionH), I get the following results:

```
call:  
lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +  
    allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +  
    cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +  
    hypertensive + influenza + injuryPoisoning + ischemicheart +  
    malignant + obesity + othercirculatorysystem + otherrespsystem +  
    renalfailure + conditionrespiratoryarrest + sepsis + vascular +  
    Northeast + South + West + allothcondsandcauses * `age45-54` +  
    covid19 * `age45-54` + influenza * `age45-54` + Northeast *  
    `age45-54` + South * `age45-54` + West * `age45-54`, data = covidcleans)
```

Residuals:

| Min     | 1Q      | Median  | 3Q     | Max    |
|---------|---------|---------|--------|--------|
| -3.9259 | -0.8079 | -0.0483 | 0.7216 | 3.3494 |

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(> t )     |
|----------------------|----------|------------|---------|--------------|
| (Intercept)          | 2.49724  | 0.18367    | 13.597  | < 2e-16 ***  |
| `age45-54`           | 0.87005  | 0.15410    | 5.646   | 2.01e-08 *** |
| adultrespdistSyn     | -0.37373 | 0.21930    | -1.704  | 0.088579 .   |
| allothcondsandcauses | 0.31877  | 0.25088    | 1.271   | 0.204096     |
| alzheimer            | -3.11804 | 0.19593    | -15.914 | < 2e-16 ***  |
| cardiacArrest        | -0.92530 | 0.23607    | -3.920  | 9.32e-05 *** |
| cardiacarrhythmia    | -2.42149 | 0.22952    | -10.550 | < 2e-16 ***  |
| cerebrovascula       | -2.34813 | 0.24262    | -9.678  | < 2e-16 ***  |
| chroniclr            | -2.05614 | 0.23056    | -8.918  | < 2e-16 ***  |
| covid19              | 1.00669  | 0.25296    | 3.980   | 7.27e-05 *** |
| diabetes             | -0.56565 | 0.22148    | -2.554  | 0.010761 *   |
| heartfailure         | -2.22678 | 0.24269    | -9.175  | < 2e-16 ***  |
| hypertensive         | -0.83520 | 0.22804    | -3.663  | 0.000260 *** |
| ...                  | -----    | -----      | -----   | -----        |

```

influenza           0.3091   0.2719   1.14  0.25585
injuryPoisoning    -2.0579   0.2258  -9.12 < 2e-16 ***
ischemicheart      -2.0304   0.2363  -8.59 < 2e-16 ***
malignant          -2.3444   0.2397  -9.78 < 2e-16 ***
obesity             -0.8527   0.2268  -3.76 0.00018 ***
othercirculatorysystem -1.4859   0.2388  -6.22 6.6e-10 ***
otherrespsystem    -2.1720   0.2425  -8.96 < 2e-16 ***
renalfailure        -1.2481   0.2259  -5.53 3.9e-08 ***
conditionrespiratoryarrest -2.5912   0.2366  -10.95 < 2e-16 ***
sepsis              -0.9008   0.2256  -3.99 6.9e-05 ***
vascular            -3.0623   0.2021  -15.16 < 2e-16 ***
Northeast            0.0898   0.1561   0.58  0.56495
South                0.5445   0.1385   3.93  8.9e-05 ***
West                 -0.1008   0.1392  -0.72  0.46925
`age45-54` :allothcondsandcauses 0.3091   0.2890   1.07  0.28499
`age45-54` :covid19     -0.1648   0.2925  -0.56  0.57331
`age45-54` :influenza    0.1692   0.3144   0.54  0.59063
`age45-54` :Northeast    0.1609   0.2164   0.74  0.45716
`age45-54` :South       0.0141   0.1920   0.07  0.94138
`age45-54` :West        0.0588   0.2013   0.29  0.77032
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.26 on 1328 degrees of freedom
Multiple R-squared:  0.561, Adjusted R-squared:  0.55
F-statistic:  53 on 32 and 1328 DF,  p-value: <2e-16

```

So, stabilizing the regression line by removing those observations and also, instead of predicting the num\_deaths, using the log(num\_deaths + 1), our adjusted R-squared increased from 0.26 to 0.55.

After pruning out age45-54:south, age45-54:West, age45-54:Influenza, age45-54:covid19, age45- 54:Northeast, age45-54:allothcondsandcauses, adultrespdistSyn

My final model is:

```
## Final model
```{r}
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast + West +
South, data
= covidcleansed)
summary(model.interactionH)
```

Inspected the model with the summary() to get the following:

```

call:
lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistsyn +
    allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
    cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
    hypertensive + influenza + injuryPoisoning + ischemicheart +
    malignant + obesity + othercirculatorysystem + otherrespssystem +
    renalfailure + conditionrespiratoryarrest + sepsis + vascular +
    Northeast + West + South, data = covidcleansed)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9493 -0.8153 -0.0476  0.7278  3.3487 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.46140   0.16836 14.620 < 2e-16 ***
`age45-54`   0.93969   0.06903 13.612 < 2e-16 ***
adultrespdistsyn -0.37378   0.21901 -1.707 0.088122 .  
allothcondsandcauses 0.47433   0.20429  2.322 0.020393 *  
alzheimer    -3.11504   0.19562 -15.924 < 2e-16 ***
cardiacArrest -0.92627   0.23556 -3.932 8.85e-05 *** 
cardiacarrhythmia -2.41400   0.22901 -10.541 < 2e-16 *** 
cerebrovascula -2.33960   0.24209 -9.664 < 2e-16 *** 
chroniclr     -2.05106   0.23016 -8.911 < 2e-16 *** 
covid19        0.92668   0.20549  4.510 7.07e-06 *** 
diabetes       -0.56834   0.22116 -2.570 0.010284 *  
heartfailure   -2.22217   0.24213 -9.178 < 2e-16 *** 
hypertensive   -0.83293   0.22764 -3.659 0.000263 *** 
influenza      0.40078   0.21297  1.882 0.060071 .  
injuryPoisoning -2.05298   0.22535 -9.110 < 2e-16 *** 
ischemicheart -2.02560   0.23569 -8.594 < 2e-16 *** 
malignant      -2.33420   0.23906 -9.764 < 2e-16 *** 
obesity        -0.85045   0.22638 -3.757 0.000180 *** 
othercirculatorysystem -1.48323   0.23851 -6.219 6.69e-10 *** 
otherrespssystem -2.16480   0.24209 -8.942 < 2e-16 *** 
renalfailure   -1.24929   0.22549 -5.540 3.63e-08 *** 
conditionrespiratoryarrest -2.58580   0.23611 -10.952 < 2e-16 *** 
sepsis         -0.90159   0.22530 -4.002 6.63e-05 *** 
vascular       -3.05905   0.20170 -15.166 < 2e-16 *** 
Northeast      0.17137   0.10774  1.591 0.111949  
West           -0.07417   0.10044 -0.739 0.460336  
South          0.54817   0.09602  5.709 1.40e-08 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.26 on 1334 degrees of freedom
Multiple R-squared:  0.5598,    Adjusted R-squared:  0.5513 
F-statistic: 65.26 on 26 and 1334 DF,  p-value: < 2.2e-16

```

Interpreting the regression coefficients of the final model:

Looking at the F-test, the P-Value is less than 2e-16. The P-Value is the probability that given the null hypothesis, that all the Betas associated with the independent variables are equal to zero. We would observe the data as extreme as it is. Since the P-values are very low, so we are going to reject the null hypothesis and accept the alternative, that at least one of the Betas is not equal to 0. We don't know which Beta or they are all not equal to zero. It is not what the F-TEST tells us. This is a test of the model itself, which tells me that something

in my model is working. Adjusted R-squared is now 0.551 meaning that 55.1 percent of the variability in the number of Covid-19 deaths is explained by the model.

Looking at the individual P-Values for Allothercondsandcauses, alzheimer , cardiacArrest, cardiacarrhythmia, cerebrovascular, chroniclr, covid19, diabetes, heartfailure, hypertensive, influenza, injuryPoisoning, ischemicheart, malignant, obesity, othercirculatorysystem, otherrespsystem, renalfailure, conditionrespiratoryarrest, sepsis , vascular, Northeast, South

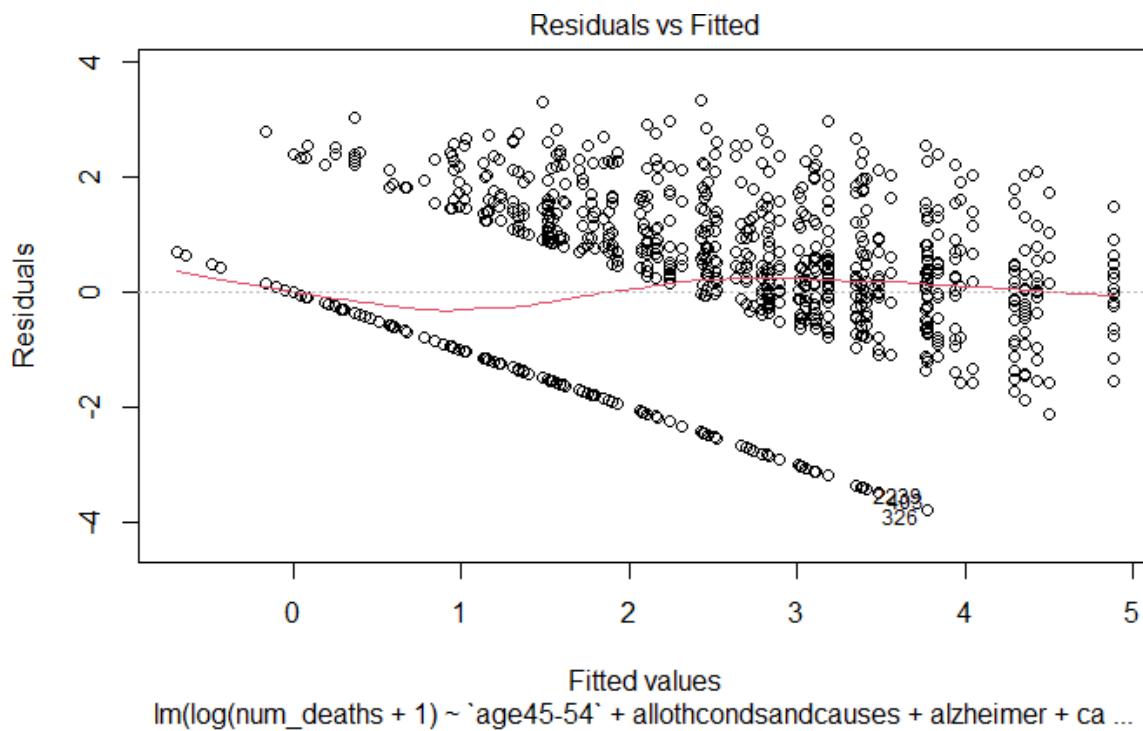
from the t-test, those are the p-values for the null hypothesis that the Betas associated with those variables are equal to zero. Because the corresponding P-Values are low, we are going to reject the null hypothesis and accept the alternative that the Betas associated with those variables are not equal to zero and then use their estimations. Remember we set dummy variables on age group with a reference level of 35-44 as and also, on conditions with reference level of Respiratoryfail so, the estimation of age35-44 will by the y-intercept of 2.46, age45-54 will be  $2.46 + 0.94 = 3.4$ , Respiratoryfail will have an estimation of 2.46, Alzheimer will have  $2.46 + (-3.11) = -0.65$ , cardiacArrest will be  $2.46 + (-0.92) = 1.58$ , cardiacarrhythmia will have an estimation of  $2.46 + (-2.414) = 0.046$ , cerebrovascular will have an estimation of  $2.46 + (-2.34) = 0.12$ , chroniclr will have an estimation of  $2.46 + (-2.05) = 0.41$ , covid19 will have an estimation of  $2.46 + 0.93 = 3.39$ , diabetes will have an estimation of  $2.46 + (-0.57) = 1.89$ , South will have an estimation of  $2.46 + 0.55 = 3.01$ , Northeast will have estimation of  $2.46 + 0.17 = 2.63$ , Midwest will have an estimation of 2.46, West will have an estimation of  $2.46 + (-0.07) = 2.39$

When we look at the worst p-values from t-tests, West and Northeast have worst p –values greater than 0.05. Recall Region variable is a dummy variable associated with MidWest, South, Northeast, and West. Because the other two have very good p-values, I will leave West and Northeast in the model since South and MidWest are doing pretty good.

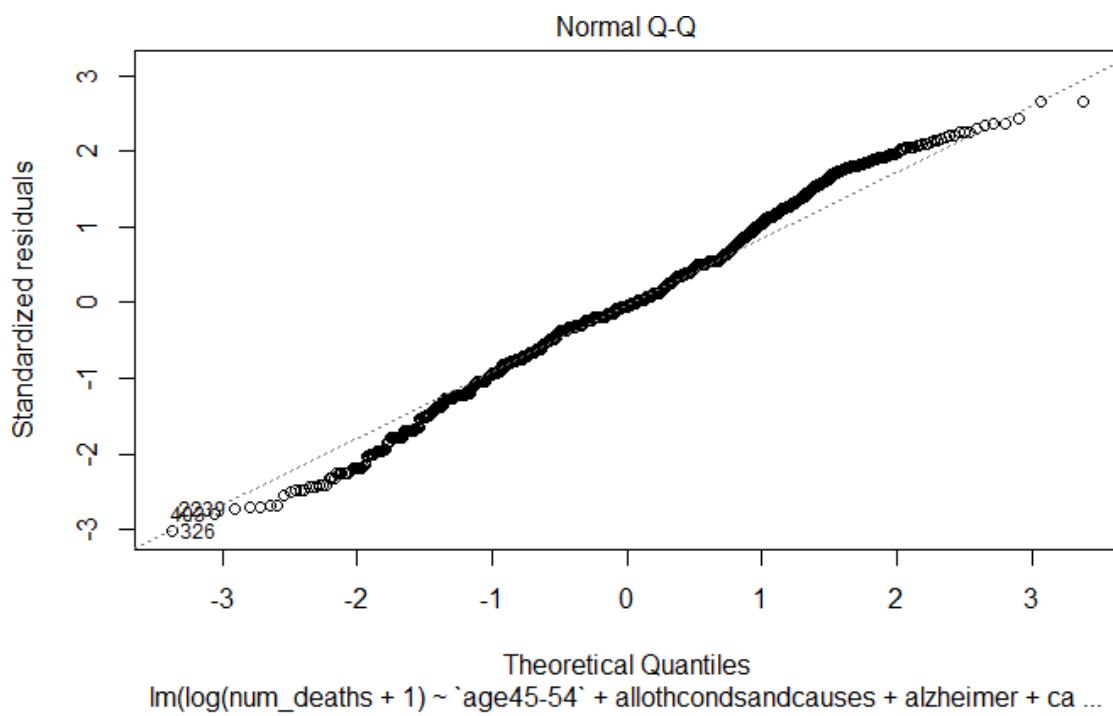
South is experiencing more covid19 deaths when compared with Northeast and MidWest. West is experiencing less deaths. Covid19 in the first place seconded by RespiratoryFailure are one of the conditions contributing to the number of Covid19 deaths.

Redoing the residual analysis by plotting the model, the following residual plots were graphed;

Looking at the first residual plot, this looks much healthier, again the fitted values against the residuals and the multiplicative trend is gone. There isn't a big change in the variance between the values on the left, middle or right.

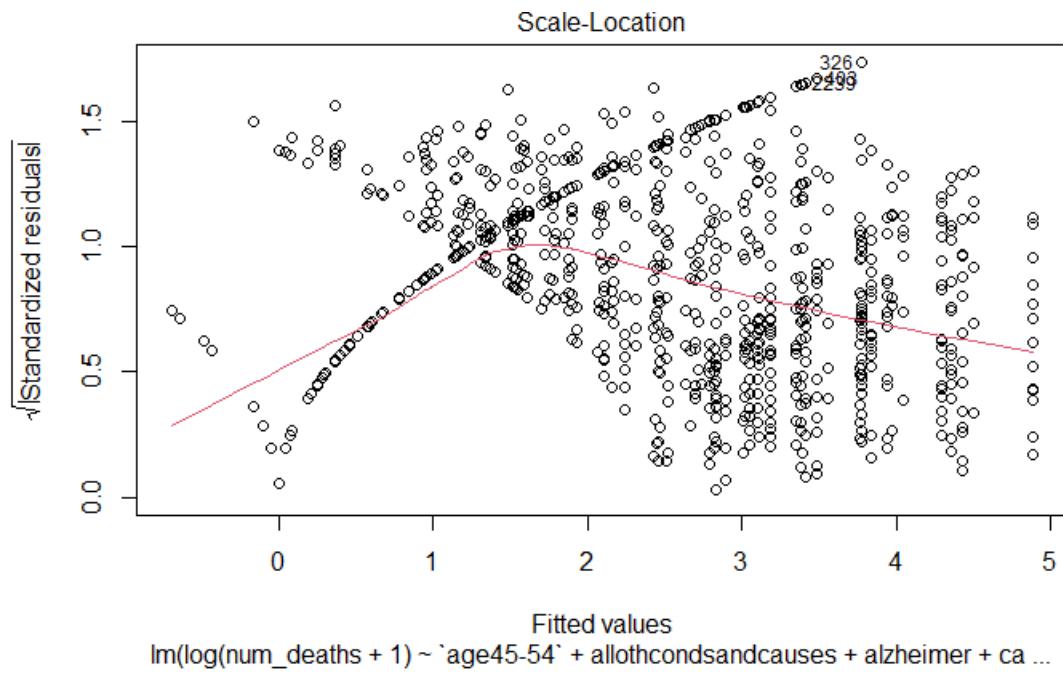


Looking at the Q-Q plot, everything is on the diagonal line. This is a normal plot which shows that the residuals are quite normal.



The third residual plot shows the fitted values against the square root of the standardized

residuals and this looks good.



The fourth residual plot showed us the leverage versus the standardized residuals. It labeled one-point, 2254, which has a lot of leverage on the regression line, and this point will have to be investigated, but right now my model looks good for now.

**Sub Section author:Bansari Gandhi**

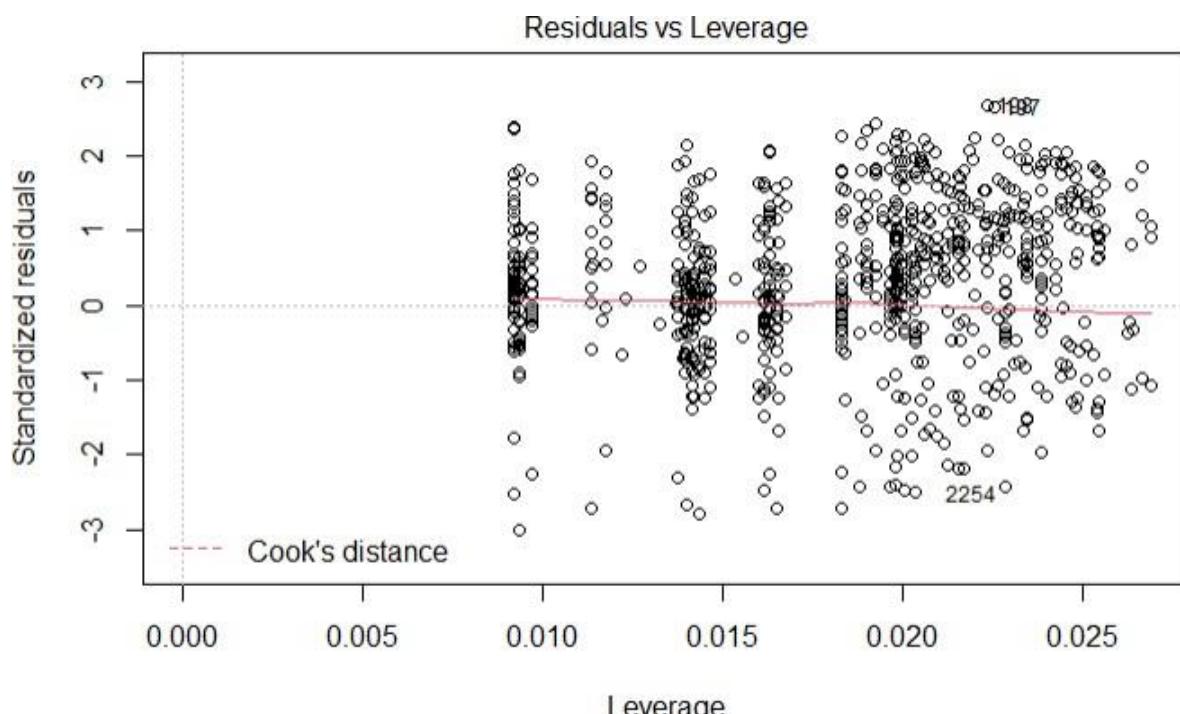
**Sub section author id: 2003383**

**Based on Module 5:**

I was specially focused on age groups that contributing to the COVID19 deaths which are 55-64 and 65-74 so I filtered the data according to my required age-group. Also, I merged dataset Covid with the state dataset to categorize state into corresponding regions. Before building the model I checked for the multicollinearity ,condition group created multicollinearity so I removed it.

First of all, I converted the age groups, Conditions, region into factors before creating dummy variables for each of the categorical variable. Then, build 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 my adjusted r-squared.

I am building first order model using N-fold cross validation using 5-folds by diving data in test and training data and predict number of COVID19 deaths with Age groups, Conditions, and regions.



I have created first order model using N-cross validation. Here is my final first order model after pruning out all the non-significant Conditions.

```
set.seed(123)

partition <- sample(2,nrow(covidcleansed),replace = TRUE, prob= c(0.80,0.20))

train <- covidcleansed[partition==1, ]

test <- covidcleansed[partition==2, ]

library(MASS)

library(DAAG)

outB <- cv.lm(data = test

    , form.lm = formula(num_deaths ~ `age65-74` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West)

    , plotit = "Observed", m=5)
```

### Analysis of Variance Table

Response: num\_deaths

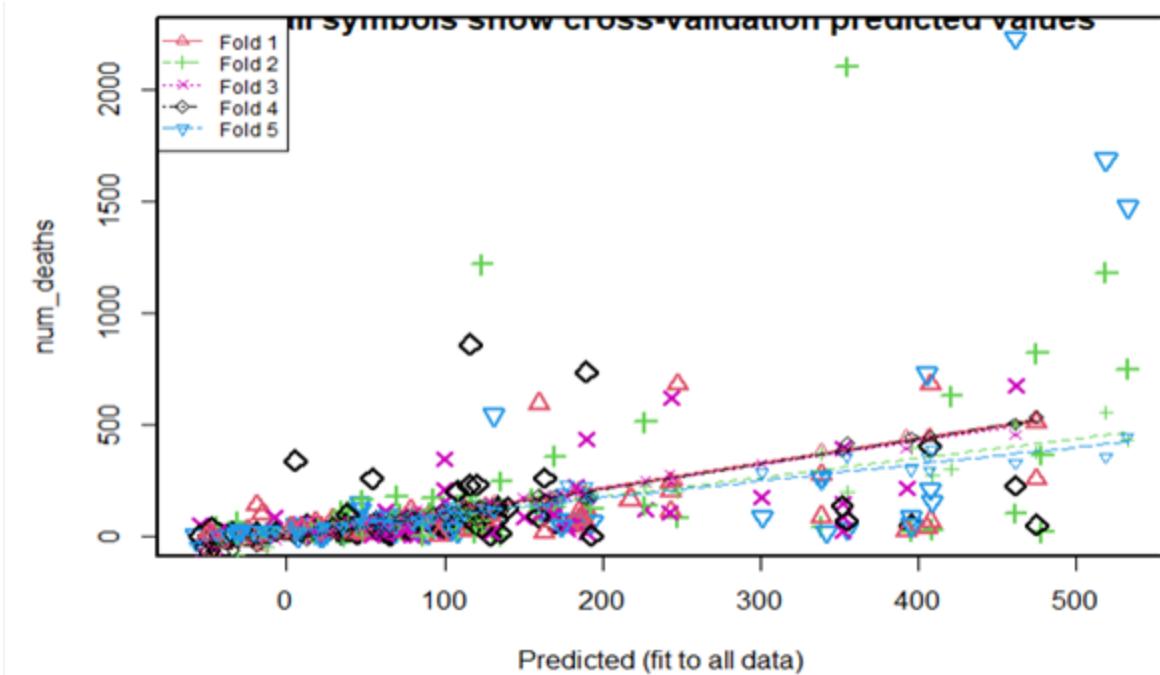
|                            | Df  | Sum Sq   | Mean Sq | F value | Pr(>F)      |
|----------------------------|-----|----------|---------|---------|-------------|
| `age65-74`                 | 1   | 346599   | 346599  | 6.41    | 0.012 *     |
| adultrespdistSyn           | 1   | 31398    | 31398   | 0.58    | 0.447       |
| allothcondsandcauses       | 1   | 1742597  | 1742597 | 32.24   | 3.2e-08 *** |
| alzheimer                  | 1   | 68407    | 68407   | 1.27    | 0.262       |
| cardiacArrest              | 1   | 21590    | 21590   | 0.40    | 0.528       |
| cardiacarrhythmia          | 1   | 45014    | 45014   | 0.83    | 0.362       |
| cerebrovascula             | 1   | 62756    | 62756   | 1.16    | 0.282       |
| chroniclr                  | 1   | 55055    | 55055   | 1.02    | 0.314       |
| covid19                    | 1   | 2084618  | 2084618 | 38.56   | 1.8e-09 *** |
| heartfailure               | 1   | 41923    | 41923   | 0.78    | 0.379       |
| hypertensive               | 1   | 39391    | 39391   | 0.73    | 0.394       |
| influenza                  | 1   | 286751   | 286751  | 5.30    | 0.022 *     |
| injuryPoisoning            | 1   | 29326    | 29326   | 0.54    | 0.462       |
| ischemicheart              | 1   | 75146    | 75146   | 1.39    | 0.239       |
| malignant                  | 1   | 7174     | 7174    | 0.13    | 0.716       |
| obesity                    | 1   | 10740    | 10740   | 0.20    | 0.656       |
| othercirculatorysystem     | 1   | 11776    | 11776   | 0.22    | 0.641       |
| otherrespystem             | 1   | 37293    | 37293   | 0.69    | 0.407       |
| renalfailure               | 1   | 2678     | 2678    | 0.05    | 0.824       |
| conditionrespiratoryarrest | 1   | 44625    | 44625   | 0.83    | 0.364       |
| sepsis                     | 1   | 8013     | 8013    | 0.15    | 0.701       |
| vascular                   | 1   | 63222    | 63222   | 1.17    | 0.280       |
| Northeast                  | 1   | 357594   | 357594  | 6.62    | 0.011 *     |
| South                      | 1   | 301060   | 301060  | 5.57    | 0.019 *     |
| West                       | 1   | 301      | 301     | 0.01    | 0.941       |
| Residuals                  | 299 | 16163293 | 54058   |         |             |

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Overall (sum over all 65 folds)

ms  
57235



```

model.firstA <- lm(num_deaths ~ `age65-74` + adultrespdistSyn + allothcondsandcauses
+
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West ,data = covidcleansed)
summary(model.firstA)

```

```

call:
lm(formula = num_deaths ~ `age65-74` + adultrespdistSyn + allothcondsandcauses +
    alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
    chroniclr + covid19 + diabetes + heartfailure + hypertensive +
    influenza + injuryPoisoning + ischemicheart + malignant +
    obesity + othercirculatorysystem + otherrespssystem + renalfailure +
    conditionrespiratoryarrest + sepsis + vascular + Northeast +
    South + West, data = covidcleansed)

Residuals:
    Min      1Q Median      3Q     Max 
-648.9  -74.3  -21.5   31.1 2928.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 157.0      29.2     5.37  9.0e-08 ***  
`age65-74`    67.0      12.2     5.49  4.7e-08 ***  
adultrespdistSyn -123.3     37.8    -3.26  0.00114 **  
allothcondsandcauses 99.9      36.8     2.71  0.00673 **  
alzheimer -221.1     43.1    -5.14  3.1e-07 ***  
cardiacArrest -152.8     39.9    -3.83  0.00013 ***  
cardiacarrhythmia -201.2     39.9    -5.05  5.0e-07 ***  
cerebrovascula -199.2     40.5    -4.91  9.8e-07 ***  
chroniclr -170.8     39.2    -4.35  1.4e-05 ***  
covid19 333.4      36.6     9.11 < 2e-16 ***  
diabetes -85.2      38.6    -2.21  0.02723 *    
heartfailure -201.2     41.3    -4.87  1.2e-06 ***  
hypertensive -84.3      38.6    -2.19  0.02897 *    
influenza 58.2      37.5     1.55  0.12116    
injuryPoisoning -212.7     40.0    -5.32  1.2e-07 ***  
ischemicheart -162.1     39.7    -4.08  4.7e-05 ***  
malignant -202.6      39.4    -5.14  3.0e-07 ***  
obesity -190.8      40.0    -4.77  2.0e-06 ***  
othercirculatorysystem -191.5     40.2    -4.77  2.1e-06 ***  
otherrespssystem -209.6     41.7    -5.03  5.4e-07 ***  
renalfailure -154.9      39.0    -3.98  7.3e-05 ***  
conditionrespiratoryarrest -217.5     46.3    -4.69  2.9e-06 ***  
sepsis -149.6      38.4    -3.89  0.00010 ***  
vascular -212.1     41.1    -5.16  2.7e-07 ***  
Northeast 102.5      19.2     5.35  1.0e-07 ***  
South 63.8       16.0     3.99  7.0e-05 ***  
West 20.6       18.4     1.12  0.26326 

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 249 on 1653 degrees of freedom
Multiple R-squared:  0.267,    Adjusted R-squared:  0.255 
F-statistic: 23.1 on 26 and 1653 DF,  p-value: <2e-16

```

This is my final first order model after pruning out variables which causes invariability in the model. When I checked for the p-value of f-test which is very low so by the null hypothesis we can reject the null hypothesis and accept the alternative that any one the beta is not equal to zero and when we check the p-value of t-test, all the values seems to be very small i.e ( $\leq 0.05$ )which is significant so by hypothesis we can reject the value and accept the alternative that we can accept the estimated value. Value of Adjusted R-

squared is 0.255 so 25.5% of variability of the number of covid deaths is explained by the model.

We can say that there were more people dying in the age group 65-74 compared to 55-64. COVID19 itself is one of the Conditions contributing to the number of COVID19 deaths. Also, most of the deaths are occurring in Northeast region compared to other regions.

Furthermore, I have created interaction model. For that, I tried different combination of specific age groups with conditions and the region variable. I tried different combinations of interaction terms and started pruning a combination one by one by seeing the value of Adjusted R-squared and checking for p-value and I did that until my adjusted R-squared got improved.

I also have created Interaction order model using N-cross validation. Here is my interaction model after pruning out all the non-significant Conditions and regions.

```
outH <- cv.lm(data = test  
 , form.lm = formula(num_deaths ~ `age65-74` + adultrespdistSyn +  
 allothcondsandcauses +  
 alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +  
 chroniclr + covid19 + diabetes + heartfailure + hypertensive +  
 influenza + injuryPoisoning + ischemicheart + malignant +  
 obesity + othercirculatorysystem + otherrespsystem + renalfailure +  
 conditionrespiratoryarrest + sepsis + vascular + Northeast +  
 South + West + allothcondsandcauses*`age65-74` + covid19*`age65-74` +  
 influenza*`age65-74` + Northeast*`age65-74` + South*`age65-74` + West*`age65-74`)
```

, plotit = "Observed", m=5)

Analysis of variance Table

Response: num\_deaths

|                                   | Df  | Sum Sq   | Mean Sq | F value | Pr(>F)  |     |
|-----------------------------------|-----|----------|---------|---------|---------|-----|
| `age65-74`                        | 1   | 346599   | 346599  | 6.88    | 0.0092  | **  |
| adultrespdistSyn                  | 1   | 31398    | 31398   | 0.62    | 0.4306  |     |
| allothcondsandcauses              | 1   | 1742597  | 1742597 | 34.57   | 1.1e-08 | *** |
| alzheimer                         | 1   | 68407    | 68407   | 1.36    | 0.2450  |     |
| cardiacArrest                     | 1   | 21590    | 21590   | 0.43    | 0.5134  |     |
| cardiacarrhythmia                 | 1   | 45014    | 45014   | 0.89    | 0.3455  |     |
| cerebrovascula                    | 1   | 62756    | 62756   | 1.24    | 0.2655  |     |
| chroniclr                         | 1   | 55055    | 55055   | 1.09    | 0.2969  |     |
| covid19                           | 1   | 2084618  | 2084618 | 41.35   | 5.2e-10 | *** |
| diabetes                          | 1   | 15       | 15      | 0.00    | 0.9862  |     |
| heartfailure                      | 1   | 42010    | 42010   | 0.83    | 0.3621  |     |
| hypertensive                      | 1   | 39390    | 39390   | 0.78    | 0.3775  |     |
| influenza                         | 1   | 291812   | 291812  | 5.79    | 0.0168  | *   |
| injuryPoisoning                   | 1   | 27194    | 27194   | 0.54    | 0.4633  |     |
| ischemicheart                     | 1   | 80362    | 80362   | 1.59    | 0.2078  |     |
| malignant                         | 1   | 5326     | 5326    | 0.11    | 0.7454  |     |
| obesity                           | 1   | 8585     | 8585    | 0.17    | 0.6802  |     |
| othercirculatorysystem            | 1   | 9879     | 9879    | 0.20    | 0.6583  |     |
| otherrespystem                    | 1   | 35077    | 35077   | 0.70    | 0.4049  |     |
| renalfailure                      | 1   | 2549     | 2549    | 0.05    | 0.8222  |     |
| conditionrespiratoryarrest        | 1   | 47643    | 47643   | 0.95    | 0.3318  |     |
| sepsis                            | 1   | 5245     | 5245    | 0.10    | 0.7473  |     |
| vascular                          | 1   | 95173    | 95173   | 1.89    | 0.1705  |     |
| Northeast                         | 1   | 371686   | 371686  | 7.37    | 0.0070  | **  |
| South                             | 1   | 313617   | 313617  | 6.22    | 0.0132  | *   |
| West                              | 1   | 450      | 450     | 0.01    | 0.9248  |     |
| `age65-74` : allothcondsandcauses | 1   | 352609   | 352609  | 6.99    | 0.0086  | **  |
| `age65-74` : covid19              | 1   | 20559    | 20559   | 0.41    | 0.5236  |     |
| `age65-74` : influenza            | 1   | 91553    | 91553   | 1.82    | 0.1788  |     |
| `age65-74` : Northeast            | 1   | 455290   | 455290  | 9.03    | 0.0029  | **  |
| `age65-74` : South                | 1   | 144052   | 144052  | 2.86    | 0.0920  | .   |
| `age65-74` : West                 | 1   | 319348   | 319348  | 6.33    | 0.0124  | *   |
| Residuals                         | 292 | 14720883 | 50414   |         |         |     |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Overall (sum over all 65 folds)

ms  
60467

^

covidcleansed\$`age65-74SQ` <- covidcleansed\$`age65-74`^2

model.interactionH <- lm(num\_deaths ~ `age65-74` + adultrespdistSyn + allothcondsandcauses +  
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +  
chroniclr + covid19 + diabetes + heartfailure + hypertensive +

```
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West +
##
allothcondsandcauses*`age65-74` + covid19*`age65-74` + influenza*`age65-74` +
Northeast*`age65-74` + South*`age65-74` + West*`age65-74`, data = covidcleansed)
summary(model.interactionH)
```

```

call:
lm(formula = num_deaths ~ `age65-74` + adultrespdistsyn + allothcondsandcauses +
alzheimer + cardiacarrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses * `age65-74` + covid19 *
`age65-74` + influenza * `age65-74` + Northeast * `age65-74` +
South * `age65-74` + West * `age65-74`, data = covidcleansed)

Residuals:
    Min      1Q Median      3Q     Max 
-757.8   -69.4  -20.6   22.9  2829.3 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           176.0     31.4     5.60  2.5e-08 ***
`age65-74`                            31.6      25.8     1.23  0.22022    
adultrespdistsyn                     -124.1     37.6    -3.30  0.00099 ***
allothcondsandcauses                  44.7      45.9     0.97  0.32974    
alzheimer                            -224.8     42.8    -5.25  1.7e-07 ***
cardiacArrest                          -151.4     39.6    -3.82  0.00014 ***
cardiacarrhythmia                     -201.8     39.6    -5.09  4.0e-07 ***
cerebrovascula                        -198.5     40.3    -4.92  9.3e-07 ***
chroniclr                             -170.0     39.0    -4.36  1.4e-05 ***
covid19                               218.8      45.3     4.83  1.5e-06 ***
diabetes                              -85.2      38.3    -2.22  0.02641 *  
heartfailure                          -199.5     41.1    -4.86  1.3e-06 ***
hypertensive                           -84.0      38.3    -2.19  0.02860 *  
influenza                             11.3      46.2     0.24  0.80675    
injuryPoisoning                       -213.0     39.8    -5.35  9.9e-08 ***
ischemicheart                         -161.8     39.5    -4.10  4.4e-05 ***
malignant                             -201.3     39.2    -5.14  3.1e-07 ***
obesity                               -191.4     39.8    -4.81  1.7e-06 ***
othercirculatorysystem                -191.8     40.0    -4.80  1.7e-06 ***
otherrespsystem                        -208.8     41.4    -5.04  5.2e-07 ***
renalfailure                           -156.2     38.8    -4.03  5.8e-05 ***
conditionrespiratoryarrest            -219.5     46.1    -4.76  2.1e-06 ***
sepsis                                -149.4     38.2    -3.91  9.7e-05 ***
vascular                             -210.4     40.9    -5.15  3.0e-07 ***
Northeast                             80.4      27.2     2.96  0.00317 ** 
South                                 53.4      22.9     2.33  0.01987 *  
West                                  26.2      26.3     1.00  0.31881    
`age65-74`:allothcondsandcauses     105.1     52.7     1.99  0.04636 *  
`age65-74`:covid19                   222.0     52.1     4.26  2.2e-05 ***
`age65-74`:influenza                 93.1      54.5     1.71  0.08816 .  
`age65-74`:Northeast                  40.1      38.1     1.05  0.29350    
`age65-74`:South                      18.9      31.9     0.59  0.55312    
`age65-74`:West                      -13.3     36.6    -0.36  0.71668   

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 248 on 1647 degrees of freedom
Multiple R-squared:  0.278,    Adjusted R-squared:  0.264 
F-statistic: 19.8 on 32 and 1647 DF,  p-value: <2e-16

```

This is my final interaction model after pruning out variables which causes invariability in

the model. When I checked for the p-value of f-test which is very low so by the null hypothesis we can reject the null hypothesis and accept the alternative that any one the beta is not equal to zero and when we check the p-value of t-test, first I check for interaction term all the values seems to be very small i.e ( $<=0.05$ )which is significant , one which are involved in interaction will going to be mandatory in the model though p-value is significant or insignificant( $<=0.05$ )so by hypothesis we can reject the value and accept the alternative that we can accept the estimated value. Value of Adjusted R-squared is 0.264 so 26.4% of variability of the number of covid deaths is explained by the model.

We can say that there were more people dying in the age group 65-74 compared to 55-64. COVID19 itself is one of the Conditions contributing to the number of COVID19 deaths. Most of the deaths are occurring in Northeast region compared to other regions. Also people who belong to age-group 65-74 and suffered from itself Covid19 contributes to the more number of covid19 deaths and people belong to age-group 65-74 and lives in north-east regions contributed highest to number of covid19 deaths.

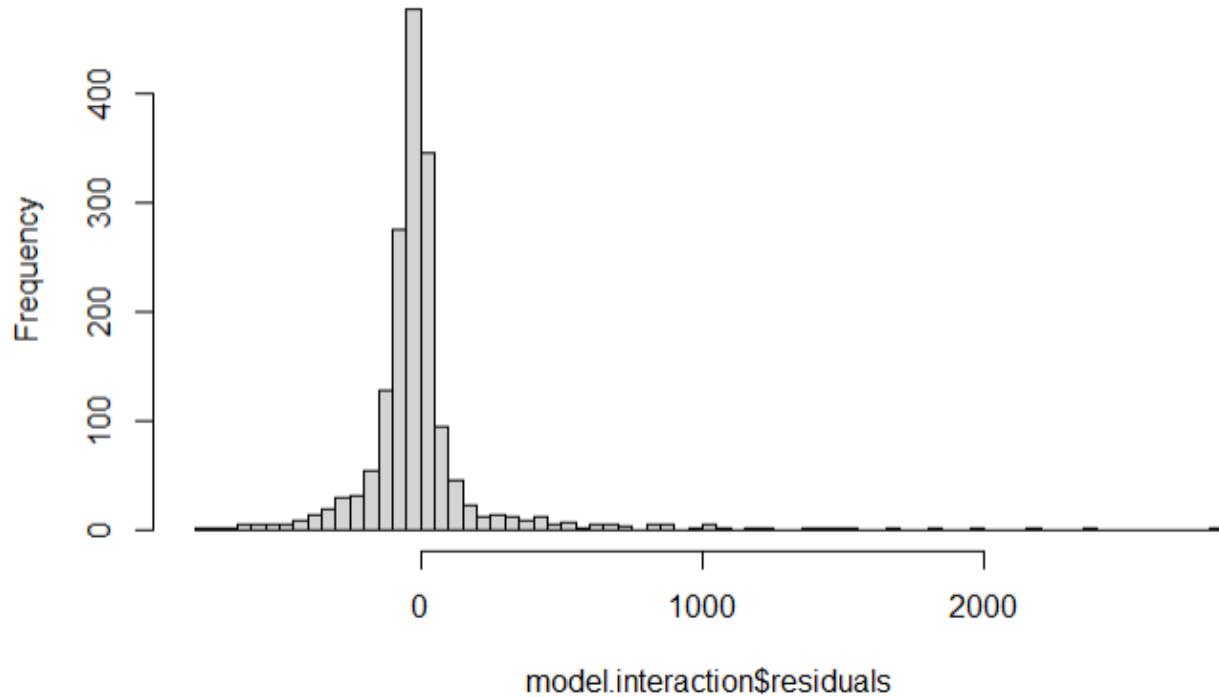
Then using interaction model I evaluated four assumptions of the residuals and did residuals

```
model.interactionH <- lm(num_deaths ~ `age65-74` + adultrespdistsyn + allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses*`age65-74` + covid19*`age65-74` + influenza*`age65-74` +
+ Northeast*`age65-74` + South*`age65-74` + West*`age65-74`, data = covidcleansed)
sum(model.interactionH$residuals)
```
[1] -2.274025e-11
```

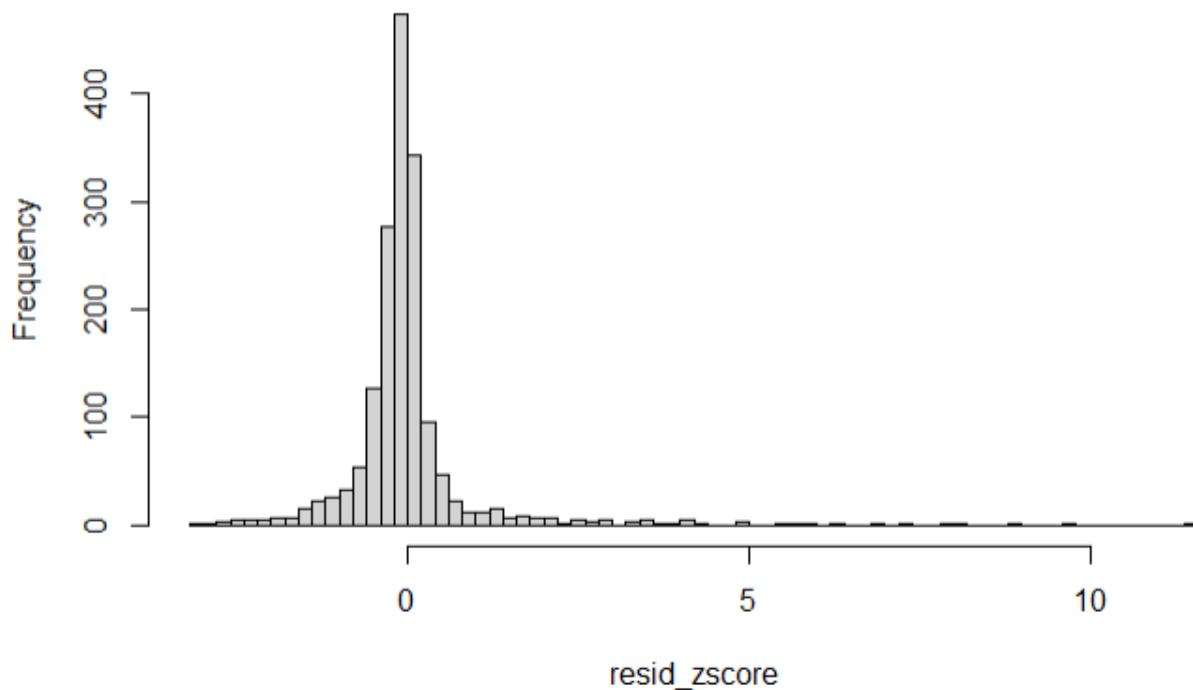
In the above histogram of model residuals, we observed data points are right skewed and there is an outlier on the right side.

Second assumption residuals are normally distributed to verified I calculate mean, standard deviation and resisd\_zscore and finally plot the histogram

**Histogram of model.interaction\$residuals**



**Histogram of resid\_zscore**



In the above histogram, we assume that 95 percent of the residuals are within two standard deviations and there are outliers on the right and left side something we need to be aware of. I used the Durbin Watson test to verify the assumption that the residuals should be independent:

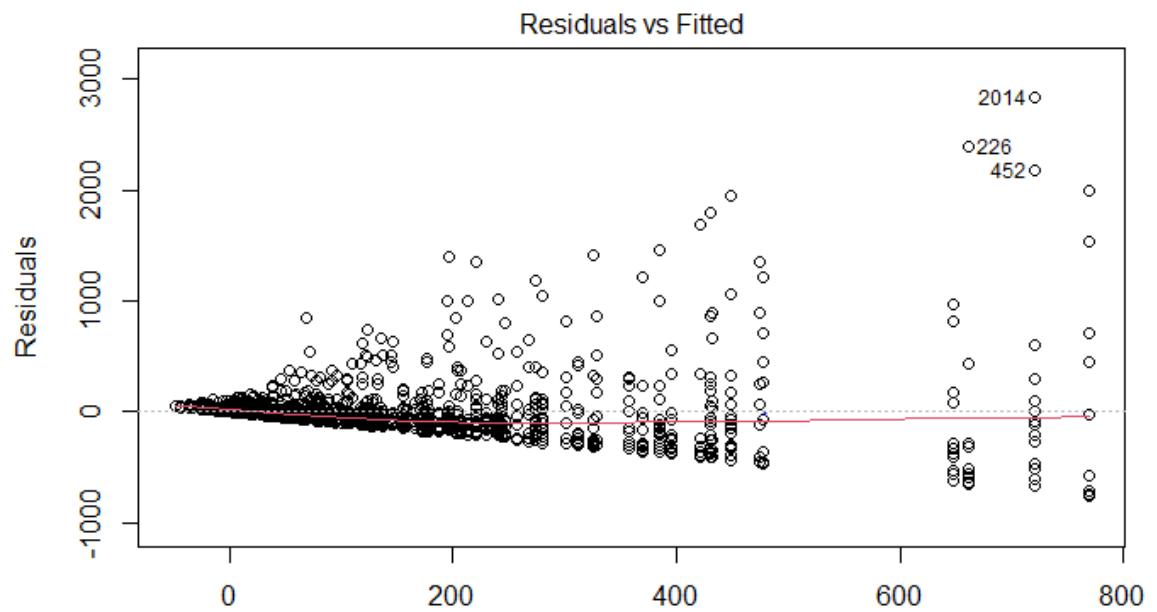
```
##verifying the independence of residuals
## using Durbin Watson test on the model.
``{r}
covidcleansed$`age65-74SQ`<- covidcleansed$`age65-74`^2
model.interaction <- lm(num_deaths ~ `age65-74` + adultrespdistsyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age65-74` + covid19*`age65-74` + influenza*`age65-74` +
  + Northeast*`age65-74` + south*`age65-74` + West*`age65-74`, data = covidcleansed)
library(car)
durbinWatsonTest(model.interaction)
```



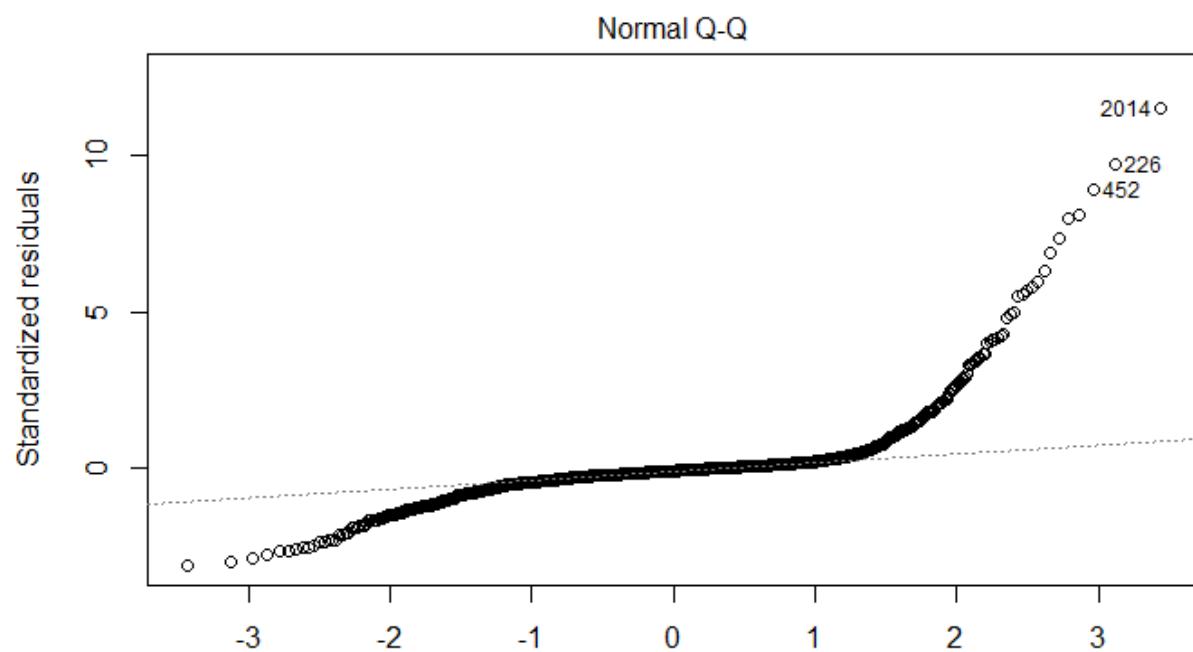
| lag | Autocorrelation | D-W Statistic | p-value |
|-----|-----------------|---------------|---------|
| 1   | 0.5950898       | 0.8097894     | 0       |


Alternative hypothesis: rho != 0

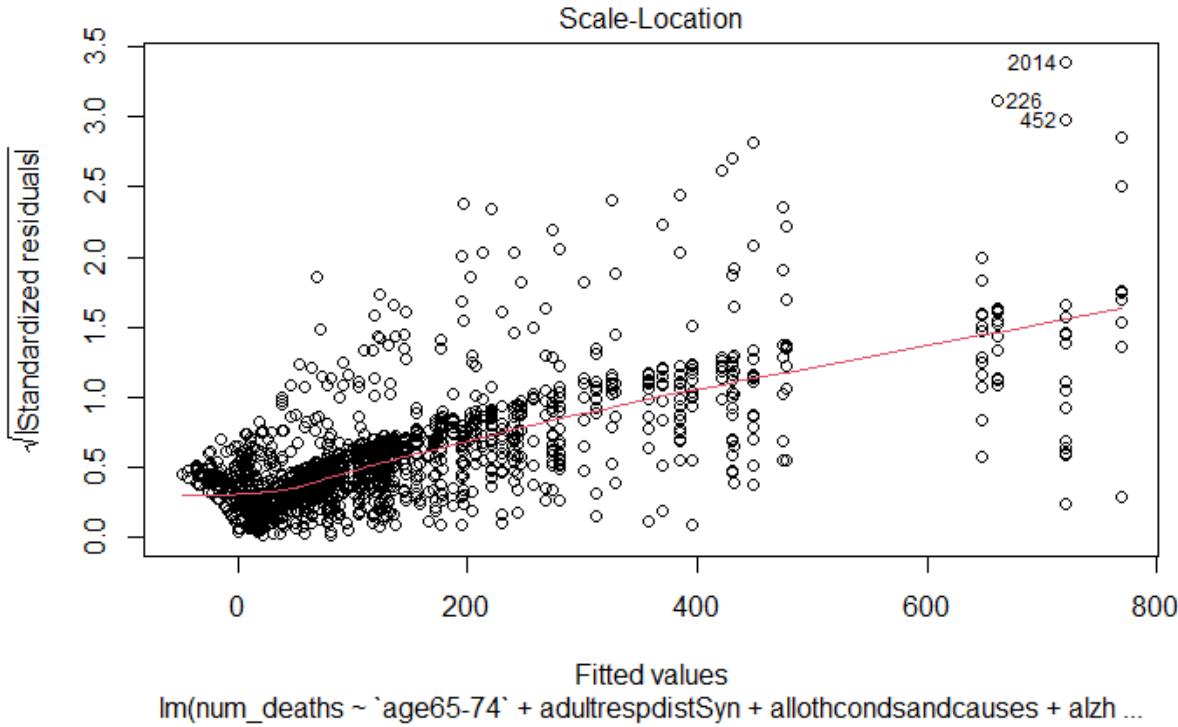
## Residual plots
``{r}
covidcleansed$`age65-74SQ`<- covidcleansed$`age65-74`^2
model.interactionH <- lm(num_deaths ~ `age65-74` + adultrespdistsyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age65-74` + covid19*`age65-74` + influenza*`age65-74` +
  + Northeast*`age65-74` + south*`age65-74` + West*`age65-74`, data = covidcleansed)
prediction <- predict(model.interaction, covidcleansed)
plot(model.interactionH)
```



`lm(num_deaths ~ `age65-74` + adultrespdistSyn + allothcondsandcauses + alzh ...`



`lm(num_deaths ~ `age65-74` + adultrespdistSyn + allothcondsandcauses + alzh ...`



```

call:
lm(formula = num_deaths ~ `age65-74` + adultrespdistSyn + allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespssystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + `age65-74`:adultrespdistsyn + `age65-74`:alzheimer +
`age65-74`:cardiacArrest + `age65-74`:cardiacarrhythmia +
`age65-74`:cerebrovascula + `age65-74`:chroniclr + `age65-74`:covid19 +
`age65-74`:diabetes + `age65-74`:heartfailure + `age65-74`:hypertensive +
`age65-74`:injuryPoisoning + `age65-74`:ischemicheart + `age65-74`:malignant +
`age65-74`:obesity + `age65-74`:othercirculatorysystem +
`age65-74`:otherrespssystem + `age65-74`:renalfailure + `age65-74`:conditionrespiratoryarrest
+
`age65-74`:sepsis + `age65-74`:vascular, data = covidcleansed)

Residuals:
    Min      1Q      Median      3Q      Max 
-743.43   -68.60    -23.53     24.92  2833.42 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 123.84     32.43   3.819 0.000139 ***
`age65-74`  132.40     30.21   4.382 1.25e-05 ***
adultrespdistSyn -77.67     48.80  -1.591 0.111692  
allothcondsandcauses 99.25     36.73   2.702 0.006960 ** 
alzheimer -162.54     53.65  -3.029 0.002490 ** 
cardiacArrest -107.48     54.23  -1.982 0.047664 *  
cardiacarrhythmia -152.97     51.61  -2.964 0.003082 ** 
cerebrovascula -147.45     53.66  -2.748 0.006065 ** 
chroniclr -128.90     52.08  -2.475 0.013428 *  
covid19  265.85     47.53   5.593 2.61e-08 *** 
diabetes -49.21     50.71  -0.970 0.332054    
heartfailure -149.04     56.22  -2.651 0.008100 ** 

```

```
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
Residual standard error: 1.29 on 1655 degrees of freedom  
Multiple R-squared: 0.4252, Adjusted R-squared: 0.4168  
F-statistic: 51 on 24 and 1655 DF, p-value: < 2.2e-16
```

This is final model made after residual analysis and Adjusted-R squared I got 0.4168 so 41.68 % of the variability of number of Covid19 death is explained by the model.

**Sub section author: Kajal Patel**

**Sub section author: 1949423**

**Based on Module 5:**

I was specially focused on age groups that contributing to the COVID19 deaths which are 75-84 and 85+. I am building first order model using N-fold cross validation and predict number of COVID19 deaths with Age groups, Conditions, Condition group and regions. Before performing to build a model, I merged the Covid dataset with the State dataset to regroup the states into corresponding regions.

First of all, I converted the age groups, Conditions, Condition groups, region into factors before creating dummy variables for each of the categorical variable. Then, build 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 my adjusted r-squared.

I have created first order model using N-cross validation. Here is my final first order model after pruning out all the non-significant Conditions.

```
set.seed(123)  
partition <- sample(2,nrow(covidcleansed),replace = TRUE, prob= c(0.80,0.20))  
train <- covidcleansed[partition==1, ]  
test <- covidcleansed[partition==2, ]  
library(MASS)  
model.FirstorderA <- lm(num_deaths ~ `age85+` + adultrespdistSyn +  
allothcondsandcauses +  
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +  
chroniclr + covid19 + diabetes + heartfailure + hypertensive +  
influenza + injuryPoisoning + ischemicheart + malignant +
```

```

obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West, data=train)

summary(model.covid19_deathsA)

prediction <- predict(model.FirstorderA,test)

actual =test$num_deaths

cor(prediction, actual)

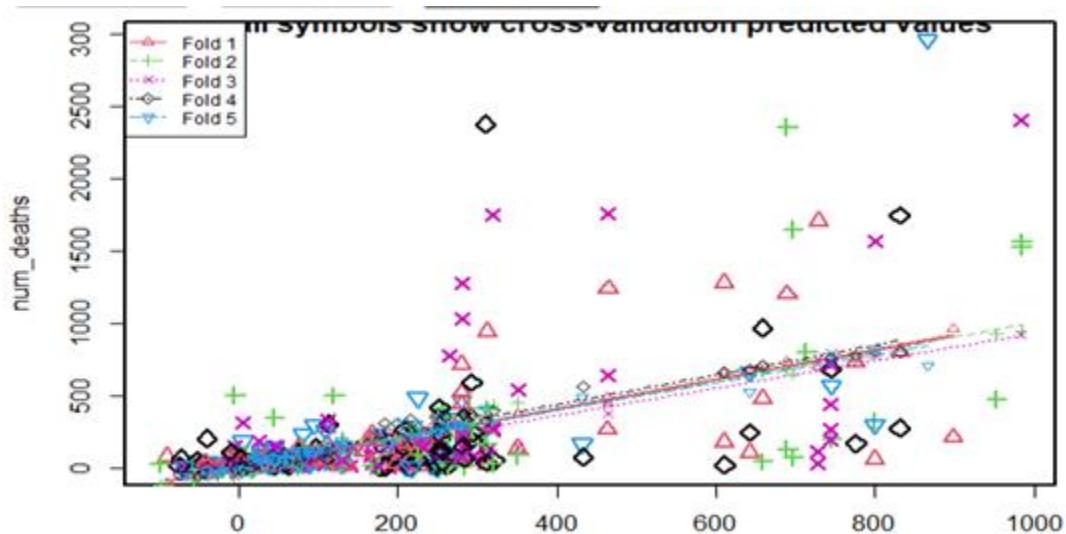
library(DAAG)

outA <- cv.lm(data = test

    , form.lm = formula(num_deaths ~ `age85+` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West)

, plotit = "Observed", m=5)

```



```

## final firstorder model
## adjusted - 0.288
``{r}
model.FirstorderA <- lm(num_deaths ~ `age85+` + adultrespdistSyn + allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West, data = covidcleansed)
summary(model.FirstorderA)
```

```

Call:  
`lm(formula = num_deaths ~ `age85+` + adultrespdistSyn + allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula + chroniclr + covid19 + diabetes + heartfailure + hypertensive + influenza + injuryPoisoning + ischemicheart + malignant + obesity + othercirculatorysystem + otherrespsystem + renalfailure + conditionrespiratoryarrest + sepsis + vascular + Northeast + South + West, data = covidcleansed)`

Residuals:

| Min   | 1Q   | Median | 3Q | Max  |
|-------|------|--------|----|------|
| -1123 | -110 | -27    | 45 | 3665 |

|                            | Estimate | Std. Error | t value | Pr(> t ) |     |
|----------------------------|----------|------------|---------|----------|-----|
| (Intercept)                | 283.806  | 41.926     | 6.77    | 1.8e-11  | *** |
| `age85+`                   | 22.862   | 17.266     | 1.32    | 0.18565  |     |
| adultrespdistSyn           | -239.031 | 56.346     | -4.24   | 2.3e-05  | *** |
| allothcondsandcauses       | 137.685  | 53.409     | 2.58    | 0.01002  | *   |
| alzheimer                  | -286.616 | 55.420     | -5.17   | 2.6e-07  | *** |
| cardiacArrest              | -232.230 | 56.960     | -4.08   | 4.8e-05  | *** |
| cardiacarrhythmia          | -256.319 | 56.350     | -4.55   | 5.8e-06  | *** |
| cerebrovascula             | -283.986 | 58.306     | -4.87   | 1.2e-06  | *** |
| chroniclr                  | -244.475 | 55.596     | -4.40   | 1.2e-05  | *** |
| covid19                    | 616.183  | 53.014     | 11.62   | < 2e-16  | *** |
| diabetes                   | -211.016 | 54.273     | -3.89   | 0.00010  | *** |
| heartfailure               | -255.291 | 55.965     | -4.56   | 5.4e-06  | *** |
| hypertensive               | -105.053 | 54.745     | -1.92   | 0.05515  | .   |
| influenza                  | 59.085   | 53.684     | 1.10    | 0.27122  |     |
| injuryPoisoning            | -319.630 | 58.335     | -5.48   | 4.9e-08  | *** |
| ischemicheart              | -197.474 | 55.776     | -3.54   | 0.00041  | *** |
| malignant                  | -306.348 | 57.179     | -5.36   | 9.5e-08  | *** |
| obesity                    | -351.777 | 64.382     | -5.46   | 5.3e-08  | *** |
| othercirculatorysystem     | -290.209 | 57.173     | -5.08   | 4.3e-07  | *** |
| otherrespsystem            | -312.967 | 58.080     | -5.39   | 8.1e-08  | *** |
| renalfailure               | -256.831 | 56.961     | -4.51   | 6.9e-06  | *** |
| conditionrespiratoryarrest | -332.711 | 67.683     | -4.92   | 9.7e-07  | *** |
| sepsis                     | -265.335 | 57.618     | -4.61   | 4.4e-06  | *** |
| vascular                   | -160.065 | 54.124     | -2.96   | 0.00314  | **  |
| Northeast                  | 216.828  | 26.946     | 8.05    | 1.5e-15  | *** |
| South                      | 62.839   | 22.661     | 2.77    | 0.00561  | **  |
| West                       | 0.755    | 25.795     | 0.03    | 0.97665  |     |

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*\*' 0.01 '\*\*' 0.05 '\*' 0.1 '.' 1

```
Residual standard error: 366 on 1775 degrees of freedom
Multiple R-squared:  0.299,    Adjusted R-squared:  0.288
F-statistic: 29.1 on 26 and 1775 DF,  p-value: <2e-16
```

First of all, I made a model by using all variables that I have in a dataset. By analyzing and checking the f-test value, the value is considered very low and equal to zero so we can reject the null hypothesis and accept the alternative. Any one of the T-test has not equaled to zero value. When I am checking p-value, I show a P-value less than 0.05 (typically  $\leq 0.05$ ) which is statistically significant. From above image, I can say that all the predictors have P- value less than the 0.05 so they all are significant.

Here, Adjusted R-square is 0.288 which means 28.8 per cent of variability in the number of COVID19 deaths was explained by the model. We can say that there 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 Northeast region.

Next, I have created interaction model. For that, I tried different combination of specific age groups and condition with the region variable. I tried to find out improved adjusted r-squared with some interaction terms. Finally, I found good adjusted r-squared for my model.

I also have created Interaction order model using N-cross validation. Here is my interaction model after pruning out all the non-significant Conditions.

```

## Final model with interaction terms
## Prunned`age85+`:chroniclr , `age85+`:obesity
## , `age85+`:heartfailure, `age85+`:cerebrovascula
## , ##`age85+`:otherrespsystem, `age85+`:malignant , `age85+`:injuryPoisoning,
`age85+`:cardiacarrhythmia, ##`age85+`:conditionrespiratoryarrest, `age85+`:vascular
## , `age85+`:alzheimer , `age85+`:renalfailure, `age85+`:othercirculatorysystem
```{r}
covidcleansed$`age85+SQ`<- covidcleansed$`age85+`^2
model.interactionH <- lm(num_deaths ~ `age85+` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses`age85+` + covid19`age85+` + influenza`age85+` +
  Northeast`age85+` + South`age85+` + West`age85+`, data = covidcleansed)
summary(model.interactionH)
```

```

```

Call:
lm(formula = num_deaths ~ `age85+` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses `age85+` + covid19 *
  `age85+` + influenza `age85+` + Northeast `age85+` +
  south `age85+` + west `age85+`, data = covidcleansed)

Residuals:
    Min      1Q  Median      3Q     Max 
-1255.1 -104.0   -25.2    47.7  3532.9 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 290.822   44.884   6.479 1.19e-10 ***
`age85+`      9.122   36.140   0.252 0.800763    
adultrespdistSyn -238.800   56.241  -4.246 2.29e-05 ***
allothcondsandcauses 122.434   65.739   1.862 0.062707    
alzheimer     -285.603   55.326  -5.162 2.72e-07 ***
cardiacArrest -231.798   56.855  -4.077 4.76e-05 ***
cardiacarrhythmia -255.994   56.246  -4.551 5.69e-06 ***
cerebrovascula -284.682   58.205  -4.891 1.09e-06 ***
chroniclr      -244.106   55.492  -4.399 1.15e-05 ***
covid19        524.002   64.892   8.075 1.24e-15 ***
diabetes       -210.944   54.172  -3.881 0.000108 ***


```

```

covid19           524.002   64.892   8.075 1.24e-15 ***
diabetes          -210.244  54.173  -3.881 0.000108 ***
heartfailure      -255.645  55.862  -4.576 5.06e-06 ***
hypertensive      -104.451  54.643  -1.912 0.056098 .
influenza         64.701   66.650   0.971 0.331802
injuryPoisoning   -320.000  58.228  -5.496 4.46e-08 ***
ischemicheart     -197.285  55.672  -3.544 0.000405 ***
malignant          -307.251  57.074  -5.383 8.29e-08 ***
obesity            -353.321  64.267  -5.498 4.41e-08 ***
othercirculatorysystem -290.758  57.067  -5.095 3.86e-07 ***
otherrespsystem    -313.283  57.972  -5.404 7.40e-08 ***
renalfailure       -257.523  56.865  -4.529 6.33e-06 ***
conditionrespiratoryarrest -333.749  67.559  -4.940 8.54e-07 ***
sepsis              -266.264  57.513  -4.630 3.93e-06 ***
vascular            -159.557  54.024  -2.953 0.003184 **
Northeast           169.878  38.481  4.415 1.07e-05 ***
South                77.605  32.175  2.412 0.015967 *
West                 8.051   36.861  0.218 0.827136
`-age85+`:allothcondsandcauses  31.832  76.923  0.414 0.679059
`-age85+`:covid19        186.179  75.827  2.455 0.014172 *
`-age85+`:influenza      -11.331  77.709  -0.146 0.884089
`-age85+`:Northeast      92.090  53.755  1.713 0.086862 .
`-age85+`:South          -30.450  45.221  -0.673 0.500809
`-age85+`:West           -14.001  51.500  -0.272 0.785756
```
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 365.4 on 1769 degrees of freedom
Multiple R-squared: 0.3036, Adjusted R-squared: 0.291
F-statistic: 24.1 on 32 and 1769 DF, p-value: < 2.2e-16

```

Here, Adjusted R-square is 0.291 which means 29.1 per cent of variability in the number of COVID19 deaths was explained by the model. We can say that there 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 Northeast region.

Here, I performed residual analysis.

1)

```

## Verrying the assumption that the sum of the residuals should add up to zero.
```{r}
model.interactionH <- lm(num_deaths ~ `age85+` + adultrespdistSyn + allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses*`age85+` + covid19*`age85+` + influenza*`age85+` +
+ Northeast*`age85+` + South*`age85+` + West*`age85+`, data = covidcleansed)
sum(model.interactionH$residuals)
```

```

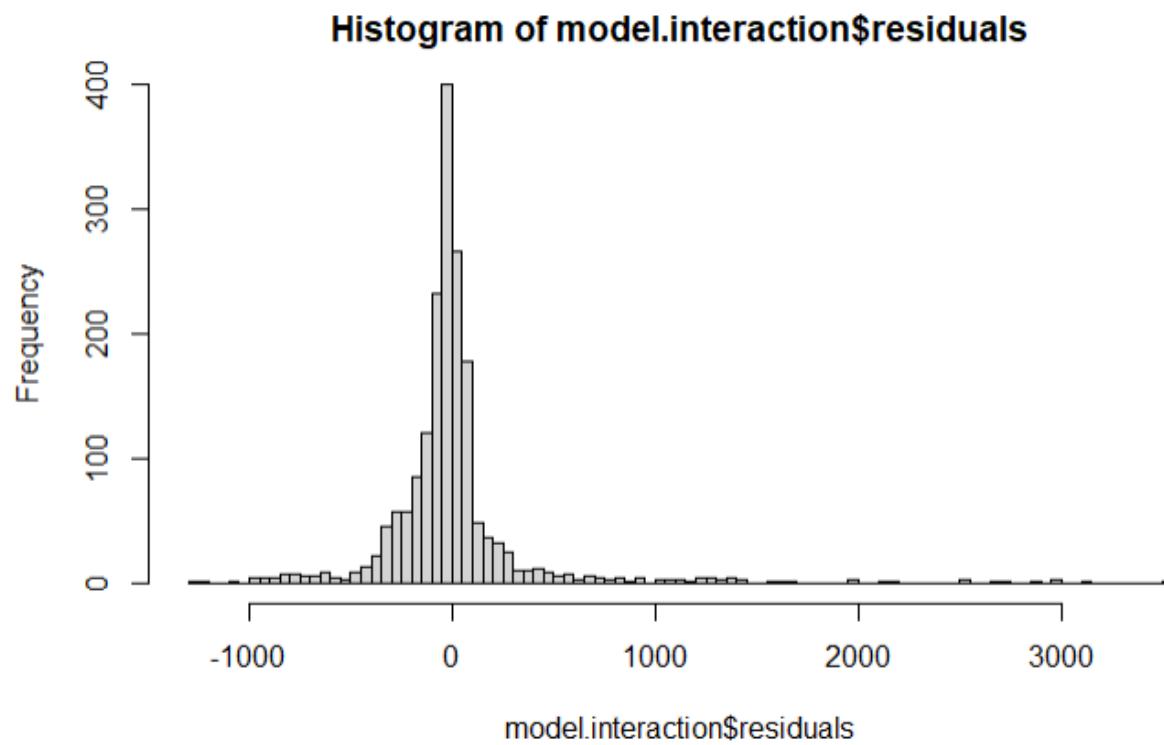
```
[1] 1.120259e-11
```

2)

```

##Histogram of residuals
```{r}
covidcleansed$`age85+`2 <- covidcleansed$`age85+`^2
model.interaction <- lm(num_deaths ~ `age85+` + adultrespdistsyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chronicTr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age85+` + covid19*`age85+` + influenza*`age85+
  + Northeast*`age85+` + South*`age85+` + West*`age85+`, data = covidcleansed)
hist(model.interaction$residuals, breaks = 100)
```

```



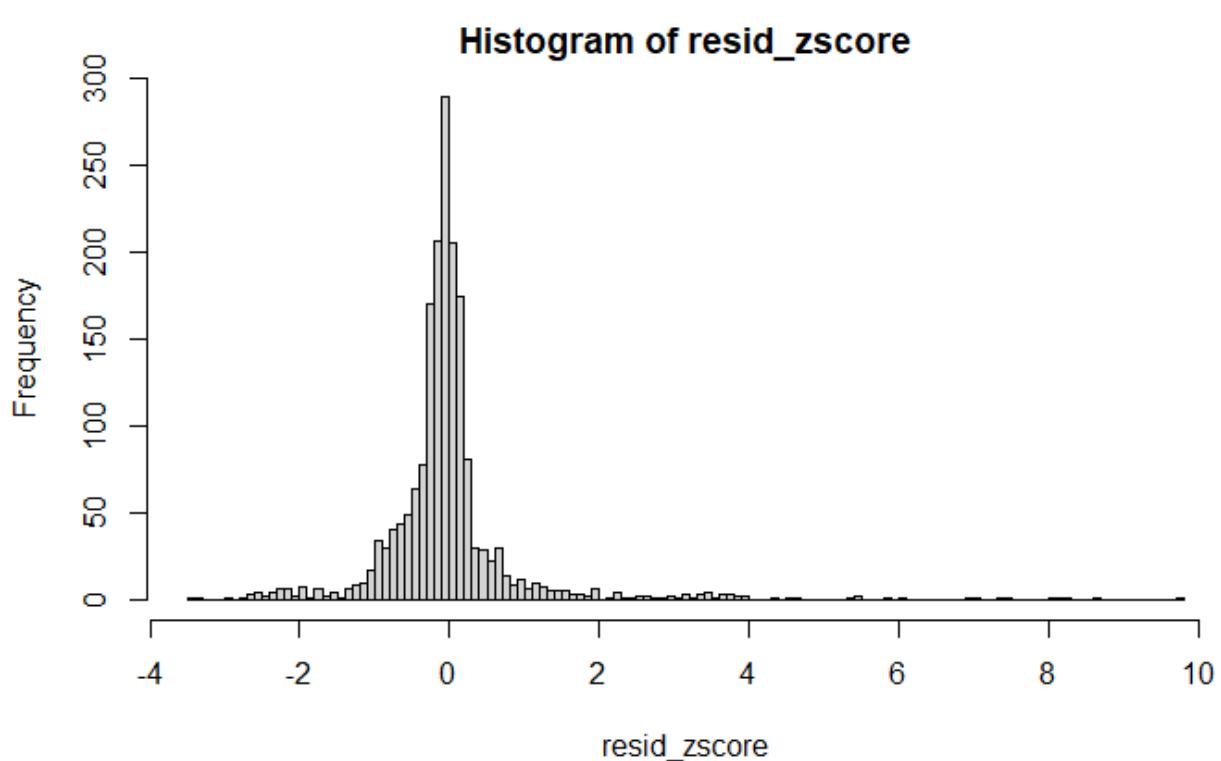
```

##idea about the distribution in the curve
```{r}
covidcleansed$`age85+SQ`<- covidcleansed$`age85+`^2
model.interaction <- lm(num_deaths ~ `age85+` + adultrespdistsyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age85+` + covid19*`age85+` + influenza*`age85+` + +
Northeast*`age85+` + South*`age85+` + West*`age85+`, data = covidcleansed)

mean = mean(model.interaction$residuals) #

## This is the standard deviation of the residuals which I am to save as sd.
sd = sd(model.interaction$residuals)
resid_zscore = (model.interaction$residuals - mean)/sd
hist(resid_zscore, breaks = 100)
```

```



3)

```

##verifying the independence of residuals
## using Durbin Watson test on the model.
````{r}
covidcleansed$`age85+SQ`<- covidcleansed$`age85+`^2
model.interaction <- lm(num_deaths ~ `age85+` + adultrespdistsyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age85+` + covid19*`age85+` + influenza*`age85+` +
  + Northeast*`age85+` + South*`age85+` + West*`age85+`, data = covidcleansed)
library(car)
durbinWatsonTest(model.interaction)
```

```

```

package car was built under R version 4.0.3
Loading required package: carData
package carData was built under R version 4.0.3
Registered S3 method overwritten by 'data.table':
  method           from
print.data.table
  lag Autocorrelation D-W Statistic p-value
  1      0.6297236    0.7405391      0
Alternative hypothesis: rho != 0

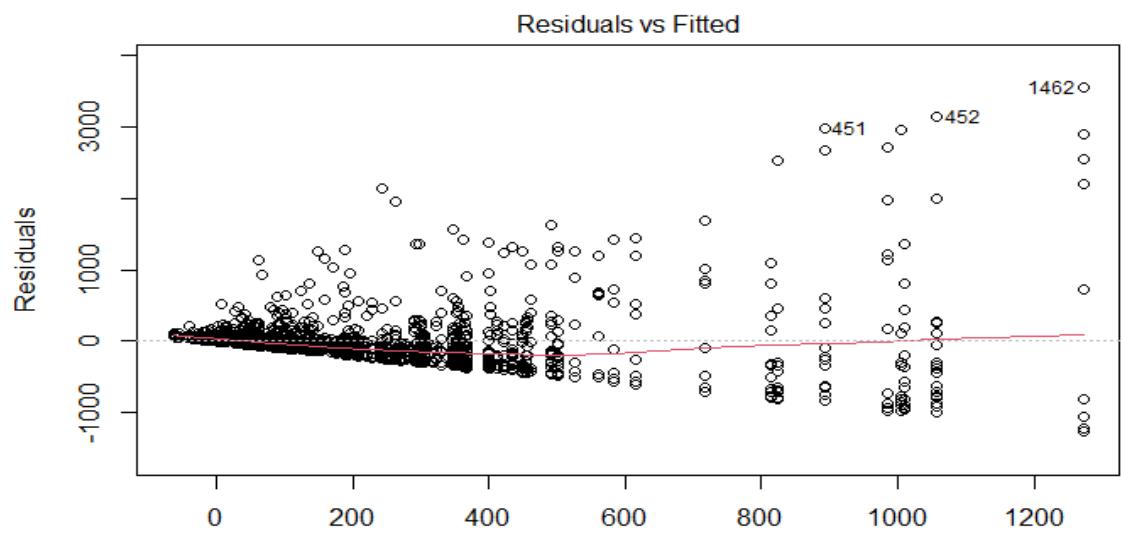
```

4)

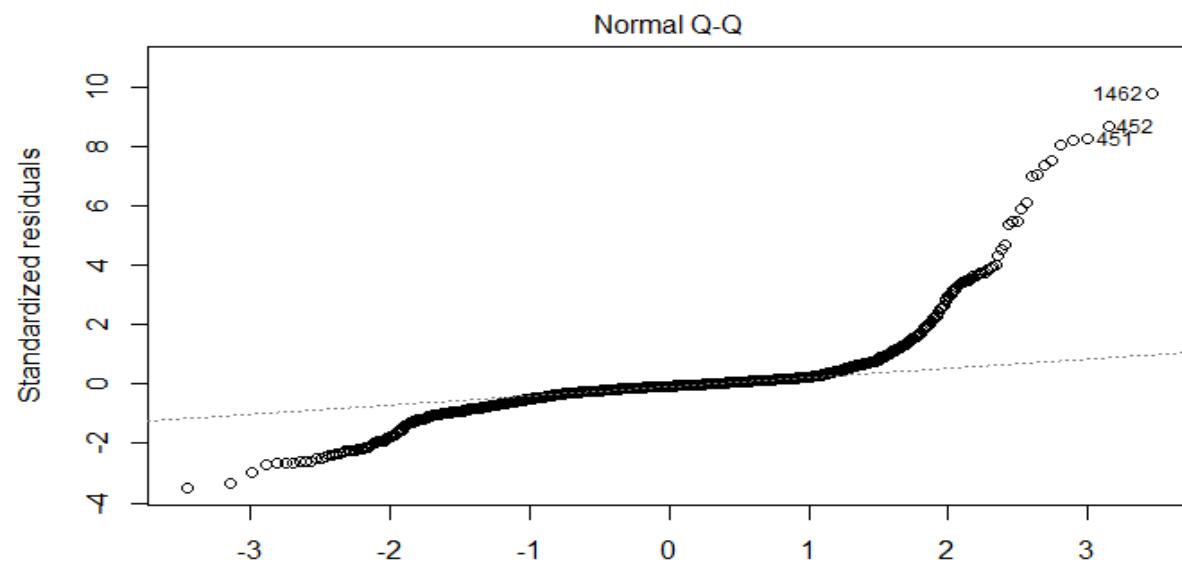
```

## Residual plots
````{r}
covidcleansed$`age85+SQ`<- covidcleansed$`age85+`^2
model.interactionH <- lm(num_deaths ~ `age85+` + adultrespdistsyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age85+` + covid19*`age85+` + influenza*`age85+` +
  + Northeast*`age85+` + south*`age85+` + West*`age85+`, data = covidcleansed)
prediction <- predict(model.interaction, covidcleansed)
plot(model.interactionH)
```

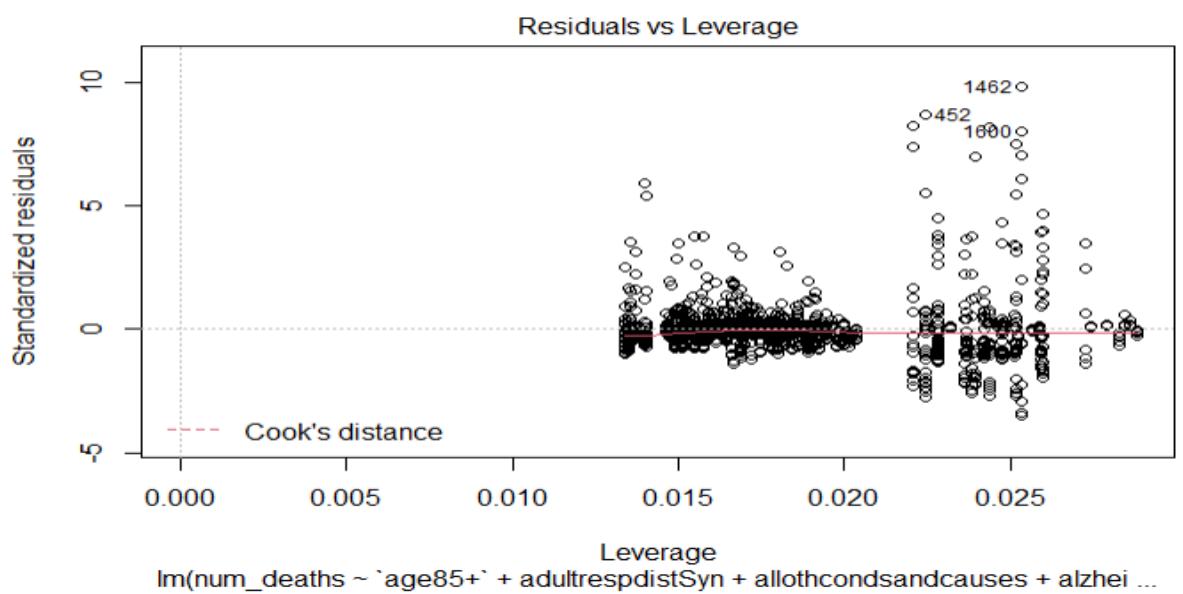
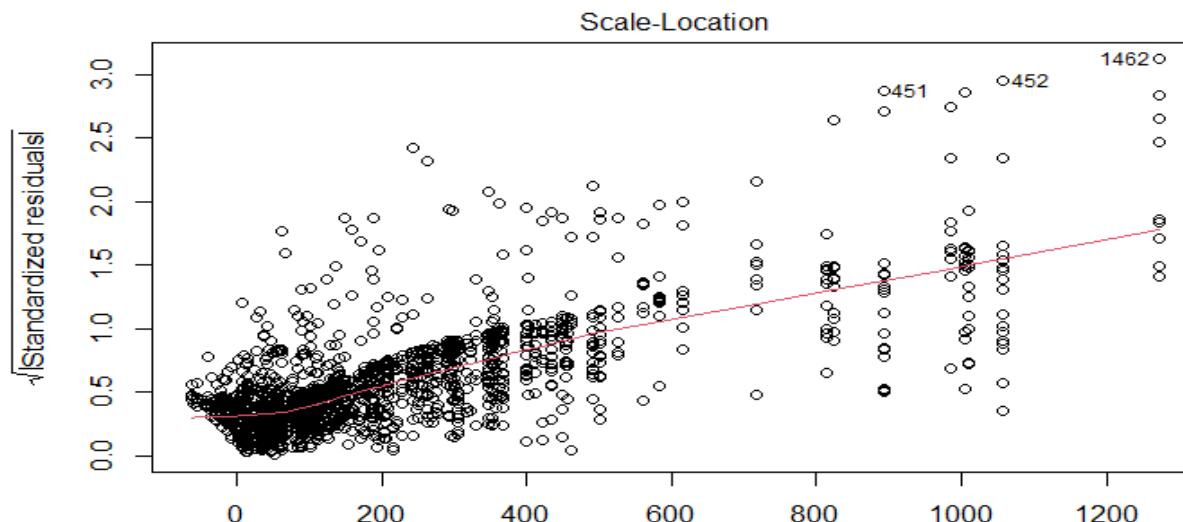
```



lm(num\_deaths ~ `age85+` + adultrespdistSyn + allothcondsandcauses + alzhei ...



lm(num\_deaths ~ `age85+` + adultrespdistSyn + allothcondsandcauses + alzhei ...



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. There are more deaths occurring in the South region compared to MidWest. Secondly, comparing the agegroup35-44 and agegroup45-54, the

agegroup45-54 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 highest 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 a highest impact on number of COVID19 deaths than other Conditions. Amongst all regions, Northeast is where people of 85+ age group are dying more.

The number of COVID19 deaths are difficult to estimate but several factors appear to be play an important role. First, COVID19 itself, respiratory failure , Influenza, all other conditions and causes are some of the conditions/ condition group contributing to the number of COVID19 deaths. Second, different age groups and regions like 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 agegroup25-34 and agegroup45-54 located in the South region are experiencing more number 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.

## Appendix

```
#title: "IndividualMilestone", Author: Ronaldlee Ejalu StudentId: 2020637
```

```
load packages
```

```
library(tidyverse)
```

```
## -- Attaching packages -----  
tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4  
## v tibble  3.0.3     v dplyr   1.0.2
```

```

## v tidyverse 1.1.2     v stringr 1.4.0
## v readr    1.3.1     v forcats 0.5.0

## -- Conflicts -----
#> #> tidyverse_conflicts() --#
#> #> dplyr::filter() masks stats::filter()
#> #> dplyr::lag()   masks stats::lag()

library(gtsummary)

## Warning: package 'gtsummary' was built under R version 4.0.3

library(tableone)

## Warning: package 'tableone' was built under R version 4.0.3

library(broom)

load data
COVID <- read.csv("C:/Users/rejalu1/OneDrive - Henry Ford Health System/DSC423/DataSets/COVID.csv")

states <- read.csv("C:/Users/rejalu1/OneDrive - Henry Ford Health System/DSC423/DataSets/states.csv")

view data
head(COVID)

## #> Data.as.of StartWeek EndWeek State Condition.Group
## #> 1 09/20/2020 02/01/2020 09/19/2020 US Respiratory diseases
## #> 2 09/20/2020 02/01/2020 09/19/2020 US Respiratory diseases
## #> 3 09/20/2020 02/01/2020 09/19/2020 US Respiratory diseases
## #> 4 09/20/2020 02/01/2020 09/19/2020 US Respiratory diseases
## #> 5 09/20/2020 02/01/2020 09/19/2020 US Respiratory diseases
## #> 6 09/20/2020 02/01/2020 09/19/2020 US Respiratory diseases
## #> Condition ICD10_codes Age.Group Number.of.COVID.19.Deaths
## #> Flag
## #> 1 Influenza and pneumonia J09-J18 0-24 142
## #> 2 Influenza and pneumonia J09-J18 25-34 655
## #> 3 Influenza and pneumonia J09-J18 35-44 1699
## #> 4 Influenza and pneumonia J09-J18 45-54 4724
## #> 5 Influenza and pneumonia J09-J18 55-64 11314
## #> 6 Influenza and pneumonia J09-J18 65-74 18305

tail(COVID)

## #> Data.as.of StartWeek EndWeek State Condition.Group Cond
## #> 12275 09/20/2020 02/16/2020 09/19/2020 YC Coronavirus Disease 2019 COV
## ID-19
## #> 12276 09/20/2020 02/17/2020 09/19/2020 YC Coronavirus Disease 2019 COV
## ID-19

```

```

## 12277 09/20/2020 02/18/2020 09/19/2020 YC Coronavirus Disease 2019 COV
ID-19
## 12278 09/20/2020 02/19/2020 09/19/2020 YC Coronavirus Disease 2019 COV
ID-19
## 12279 09/20/2020 02/20/2020 09/19/2020 YC Coronavirus Disease 2019 COV
ID-19
## 12280 09/20/2020 02/21/2020 09/19/2020 YC Coronavirus Disease 2019 COV
ID-19
##           ICD10_codes Age.Group Number.of.COVID.19.Deaths
## 12275          U071    55-64                      3283
## 12276          U071    65-74                      5039
## 12277          U071    75-84                      5407
## 12278          U071     85+                      4858
## 12279          U071 Not stated                         NA
## 12280          U071 All ages                     20707
##                                     Flag
## 12275
## 12276
## 12277
## 12278
## 12279 Counts less than 10 suppressed.
## 12280

```

### data structure

`str(COVID) # Check variables so that there are of the right class.`

```

## 'data.frame':   12280 obs. of  10 variables:
## $ Data.as.of : chr "09/20/2020" "09/20/2020" "09/20/2020"
## "09/20/2020" ...
## $ StartWeek  : chr "02/01/2020" "02/01/2020" "02/01/2020"
## "02/01/2020" ...
## $ EndWeek   : chr "09/19/2020" "09/19/2020" "09/19/2020"
## "09/19/2020" ...
## $ State     : chr "US" "US" "US" "US" ...
## $ Condition.Group : chr "Respiratory diseases" "Respiratory dis
eases" "Respiratory diseases" "Respiratory diseases" ...
## $ Condition   : chr "Influenza and pneumonia" "Influenza an
d pneumonia" "Influenza and pneumonia" "Influenza and pneumonia" ...
## $ ICD10_codes : chr "J09-J18" "J09-J18" "J09-J18" "J09-J18"
...
## $ Age.Group      : chr "0-24" "25-34" "35-44" "45-54" ...
## $ Number.of.COVID.19.Deaths: int 142 655 1699 4724 11314 18305 21231 217
25 2 79797 ...
## $ Flag          : chr " " " " " "

```

### Select variables of interest for cleaning

```

covid_select <- COVID %>%
  select(state = State
        ,condition_group = Condition.Group
        ,condition = Condition

```



```

## convert the condition model matrix into a data frame
conditiondummies.frame <- data.frame(conditiondummies.matrix)
covid_clean <- cbind(covid_clean, conditiondummies.frame)

## create a matrix for regions
regiondummies.matrix <- model.matrix(~covid_clean$region)
## convert the model matrix into a data frame
regiondummies.frame <- data.frame(regiondummies.matrix)
covid_clean <- cbind(covid_clean, regiondummies.frame)

```

Select variables of interest after data cleaning

```

covidcleansed <- covid_clean %>%
  select( num_deaths = num_deaths
    , condition = condition
    , 'age35-44' = covid_clean.age_group35.44
    , 'age45-54' = covid_clean.age_group45.54
    , 'respiratoryfail' = covid_clean.conditionRespiratoryfail
    , adultrespdistSyn = covid_clean.conditionAdultRespDistSyn
    , alloothcondsandcauses = covid_clean.conditionAll.other.conditions.a
nd.causes..residual.
    , alzheimer = covid_clean.conditionAlzheimer
    , cardiacArrest = covid_clean.conditionCardiacArrest
    , cardiacarrhythmia = covid_clean.conditionCardiacArrhythmia
    , cerebrovascula = covid_clean.conditionCerebrovascular
    , chroniclr = covid_clean.conditionChronicLR
    , covid19 = covid_clean.conditionCOVID.19
    , diabetes = covid_clean.conditionDiabetes
    , heartfailure = covid_clean.conditionHeartFailure
    , hypertensive = covid_clean.conditionHypertensive
    , influenza = covid_clean.conditionInfluenza
    , injuryPoisoning = covid_clean.conditioninjuryPoisoning
    , ischemicheart = covid_clean.conditionIschemicheart
    , malignant = covid_clean.conditionMalignant
    , obesity = covid_clean.conditionObesity
    , othercirculatorysystem = covid_clean.conditionOtherCirculatorySyst
em
    , otherrespystem = covid_clean.conditionOtherRespSystem
    , renalfailure = covid_clean.conditionRenal.failure
    , conditionrespiratoryarrest = covid_clean.conditionRespiratoryArres
t
    , sepsis = covid_clean.conditionSepsis
    , vascular = covid_clean.conditionVascular
    , Midwest = covid_clean.regionMidwest
    , Northeast = covid_clean.regionNortheast
    , South = covid_clean.regionSouth
    , West = covid_clean.regionWest
  )
str(covidcleansed)

```

```

## 'data.frame': 1364 obs. of 27 variables:
## $ num_deaths : int 0 0 0 0 0 0 0 0 0 0 ...
## $ age45-54   : num 0 0 1 0 0 0 0 1 0 0 ...
## $ adultrespdistSyn : num 0 0 0 1 0 0 0 0 0 0 ...
## $ allothcondsandcauses : num 0 0 0 0 0 0 0 0 0 0 ...
## $ alzheimer  : num 0 0 0 0 0 0 0 0 0 0 ...
## $ cardiacArrest : num 0 0 0 0 0 0 0 0 0 1 ...
## $ cardiacarrhythmia : num 0 0 0 0 0 0 0 0 0 0 ...
## $ cerebrovascula : num 0 0 0 0 0 0 0 0 0 0 ...
## $ chroniclr   : num 0 1 1 0 0 0 0 0 0 0 ...
## $ covid19     : num 0 0 0 0 0 0 0 0 0 0 ...
## $ diabetes    : num 0 0 0 0 0 0 0 0 0 0 ...
## $ heartfailure: num 0 0 0 0 0 0 0 0 0 0 ...
## $ hypertensive: num 0 0 0 0 0 0 1 1 0 0 ...
## $ influenza   : num 1 0 0 0 0 0 0 0 0 0 ...
## $ injuryPoisoning: num 0 0 0 0 0 0 0 0 0 0 ...
## $ ischemicheart: num 0 0 0 0 0 0 0 0 0 1 0 ...
## $ malignant   : num 0 0 0 0 0 0 0 0 0 0 ...
## $ obesity     : num 0 0 0 0 0 0 0 0 0 0 ...
## $ othercirculatorysystem: num 0 0 0 0 0 0 0 0 0 0 ...
## $ otherrespystem: num 0 0 0 0 0 1 0 0 0 0 ...
## $ renalfailure: num 0 0 0 0 0 0 0 0 0 0 ...
## $ conditionrespiratoryarrest: num 0 0 0 0 0 0 0 0 0 0 ...
## $ sepsis      : num 0 0 0 0 0 0 0 0 0 0 ...
## $ vascular   : num 0 0 0 0 0 0 0 0 0 0 ...
## $ Northeast   : num 0 0 0 0 0 0 0 0 0 0 ...
## $ South       : num 0 0 0 0 0 0 0 0 0 0 ...
## $ West        : num 1 1 1 1 1 1 1 1 1 1 ...

```

[summary](#)(covidcleansed)

|              | num_deaths | age45-54        | adultrespdistSyn  | allothcondsandcauses |
|--------------|------------|-----------------|-------------------|----------------------|
| ## Min.      | : 0.00     | Min. :0.0000    | Min. :0.00000     | Min. :0.00000        |
| ## 1st Qu.:  | 0.00       | 1st Qu.:0.0000  | 1st Qu.:0.00000   | 1st Qu.:0.00000      |
| ## Median :  | 10.00      | Median :1.0000  | Median :0.00000   | Median :0.00000      |
| ## Mean :    | 31.38      | Mean :0.5073    | Mean :0.04545     | Mean :0.06012        |
| ## 3rd Qu.:  | 27.00      | 3rd Qu.:1.0000  | 3rd Qu.:0.00000   | 3rd Qu.:0.00000      |
| ## Max. :    | 1185.00    | Max. :1.0000    | Max. :1.00000     | Max. :1.00000        |
| ## alzheimer |            | cardiacArrest   | cardiacarrhythmia | cerebrovascula       |
| ## Min. :    | 0.00000    | Min. :0.00000   | Min. :0.00000     | Min. :0.00000        |
| ## 1st Qu.:  | 0.00000    | 1st Qu.:0.00000 | 1st Qu.:0.00000   | 1st Qu.:0.00000      |
| ## Median :  | 0.00000    | Median :0.00000 | Median :0.00000   | Median :0.00000      |
| ## Mean :    | 0.07331    | Mean :0.03519   | Mean :0.03886     | Mean :0.03226        |
| ## 3rd Qu.:  | 0.00000    | 3rd Qu.:0.00000 | 3rd Qu.:0.00000   | 3rd Qu.:0.00000      |
| ## Max. :    | 1.00000    | Max. :1.00000   | Max. :1.00000     | Max. :1.00000        |
| ## chroniclr |            | covid19         | diabetes          | heartfailure         |
| ## Min. :    | 0.00000    | Min. :0.00000   | Min. :0.00000     | Min. :0.00000        |
| ## 1st Qu.:  | 0.00000    | 1st Qu.:0.00000 | 1st Qu.:0.00000   | 1st Qu.:0.00000      |
| ## Median :  | 0.00000    | Median :0.00000 | Median :0.00000   | Median :0.00000      |
| ## Mean :    | 0.03812    | Mean :0.06085   | Mean :0.04399     | Mean :0.03226        |

```

## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000
## Max. :1.00000 Max. :1.00000 Max. :1.00000 Max. :1.00000
## hypertensive influenza injuryPoisoning ischemicheart
## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000
## Median :0.00000 Median :0.00000 Median :0.00000 Median :0.00000
## Mean :0.03959 Mean :0.05059 Mean :0.04106 Mean :0.03519
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000
## Max. :1.00000 Max. :1.00000 Max. :1.00000 Max. :1.00000
## malignant obesity othercirculatorysystem otherrespsyste
m
## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000
## Median :0.00000 Median :0.00000 Median :0.00000 Median :0.00000
## Mean :0.03372 Mean :0.04032 Mean :0.03372 Mean :0.0322
6
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000
## Max. :1.00000 Max. :1.00000 Max. :1.00000 Max. :1.00000
## renalfailure conditionrespiratoryarrest sepsis
## Min. :0.00000 Min. :0.00000 Min. :0.00000
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000
## Median :0.00000 Median :0.00000 Median :0.00000
## Mean :0.04106 Mean :0.03519 Mean :0.04106
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000
## Max. :1.00000 Max. :1.00000 Max. :1.00000
## vascular Northeast South West
## Min. :0.00000 Min. :0.00000 Min. :0.00000 Min. :0.00000
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000
## Median :0.00000 Median :0.00000 Median :0.00000 Median :0.00000
## Mean :0.06378 Mean :0.1957 Mean :0.3306 Mean :0.2639
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :1.00000 Max. :1.00000 Max. :1.00000 Max. :1.00000

```

##run the correlation table

```

cor(covidcleansed)

##                               num_deaths    age45-54 adultrespdistSyn
## num_deaths           1.0000000000  0.189203611   0.01228337
## age45-54            0.1892036112  1.000000000   0.01088007
## adultrespdistSyn   0.0122833701  0.010880067   1.000000000
## allothcondsandcauses 0.1456681604 -0.003708730  -0.05518909
## alzheimer          -0.1098532943 -0.009750710  -0.06137858
## cardiacArrest       0.0002345289  0.021074705  -0.04167571
## cardiacarrhythmia -0.0698400501 -0.029507102  -0.04387599

```

|                               |               |               |              |
|-------------------------------|---------------|---------------|--------------|
| ## cerebrovascula             | -0.0607191227 | -0.035876204  | -0.03984095  |
| ## chroniclr                  | -0.0584080657 | -0.002919423  | -0.04344353  |
| ## covid19                    | 0.3461343754  | 0.005468677   | -0.05554626  |
| ## diabetes                   | 0.0209647686  | 0.054061837   | -0.04680876  |
| ## heartfailure               | -0.0606158362 | -0.044175923  | -0.03984095  |
| ## hypertensive               | 0.0078358731  | 0.027104516   | -0.04430488  |
| ## influenza                  | 0.1389282132  | 0.013343562   | -0.05037094  |
| ## injuryPoisoning            | -0.0627855216 | -0.003034258  | -0.04515236  |
| ## ischemicheart              | -0.0549304442 | -0.034634398  | -0.04167571  |
| ## malignant                  | -0.0627534941 | -0.018986460  | -0.04076726  |
| ## obesity                    | -0.0044959582 | 0.023085275   | -0.04473031  |
| ## othercirculatorysystem     | -0.0437480506 | -0.002739577  | -0.04076726  |
| ## otherrespsystem            | -0.0573623121 | -0.019276766  | -0.03984095  |
| ## renalfailure               | -0.0271461038 | 0.063481005   | -0.04515236  |
| ## conditionrespiratoryarrest | -0.0689445262 | -0.010759068  | -0.04167571  |
| ## sepsis                     | -0.0149137360 | 0.011746912   | -0.04515236  |
| ## vascular                   | -0.1019414484 | -0.036832946  | -0.05695797  |
| ## Northeast                  | 0.0105189052  | 0.031572127   | -0.00120963  |
| ## South                      | 0.0831565833  | 0.044242943   | 0.01870378   |
| ## West                       | -0.0298516097 | -0.071994896  | -0.00290377  |
| ## allothcondsandcauses       |               | alzheimer     | cardiacArres |
| t                             |               |               |              |
| ## num_deaths                 | 0.145668160   | -0.1098532943 | 0.000234528  |
| 9                             |               |               |              |
| ## age45-54                   | -0.003708730  | -0.0097507103 | 0.021074704  |
| 5                             |               |               |              |
| ## adultrespdistSyn           | -0.055189094  | -0.0613785770 | -0.041675711 |
| 9                             |               |               |              |
| ## allothcondsandcauses       | 1.000000000   | -0.0711359895 | -0.048300940 |
| 5                             |               |               |              |
| ## alzheimer                  | -0.071135989  | 1.0000000000  | -0.053717913 |
| 7                             |               |               |              |
| ## cardiacArrest              | -0.048300941  | -0.0537179137 | 1.000000000  |
| 0                             |               |               |              |
| ## cardiacarrhythmia          | -0.050851001  | -0.0565539641 | -0.038399826 |
| 6                             |               |               |              |
| ## cerebrovascula             | -0.046174509  | -0.0513530019 | -0.034868402 |
| 2                             |               |               |              |
| ## chroniclr                  | -0.050349791  | -0.0559965431 | -0.038021340 |
| 8                             |               |               |              |
| ## covid19                    | -0.064376505  | -0.0715963602 | -0.048613529 |
| 7                             |               |               |              |
| ## diabetes                   | -0.054250000  | -0.0603341630 | -0.040966560 |
| 6                             |               |               |              |
| ## heartfailure               | -0.046174509  | -0.0513530019 | -0.034868402 |
| 2                             |               |               |              |
| ## hypertensive               | -0.051348072  | -0.0571067829 | -0.038775187 |
| 5                             |               |               |              |
| ## influenza                  | -0.058378452  | -0.0649256228 | -0.044084136 |
| 2                             |               |               |              |

|                               |                   |                |              |
|-------------------------------|-------------------|----------------|--------------|
| ## injuryPoisoning            | -0.052330280      | -0.0581991450  | -0.039516895 |
| 2                             |                   |                |              |
| ## ischemicheart              | -0.048300941      | -0.0537179137  | -0.036474164 |
| 1                             |                   |                |              |
| ## malignant                  | -0.047248076      | -0.0525469694  | -0.035679099 |
| 5                             |                   |                |              |
| ## obesity                    | -0.051841127      | -0.0576551338  | -0.039147514 |
| 7                             |                   |                |              |
| ## othercirculatorysystem     | -0.047248076      | -0.0525469694  | -0.035679099 |
| 5                             |                   |                |              |
| ## otherrespystem             | -0.046174509      | -0.0513530019  | -0.034868402 |
| 2                             |                   |                |              |
| ## renalfailure               | -0.052330280      | -0.0581991450  | -0.039516895 |
| 2                             |                   |                |              |
| ## conditionrespiratoryarrest | -0.048300941      | -0.0537179137  | -0.036474164 |
| 1                             |                   |                |              |
| ## sepsis                     | -0.052330280      | -0.0581991450  | -0.039516895 |
| 2                             |                   |                |              |
| ## vascular                   | -0.066012638      | -0.0734159868  | -0.049849046 |
| 0                             |                   |                |              |
| ## Northeast                  | -0.008172205      | -0.0111635365  | 0.006057866  |
| 6                             |                   |                |              |
| ## South                      | 0.018927842       | -0.0003857318  | 0.001091290  |
| 7                             |                   |                |              |
| ## West                       | -0.025486637      | -0.0088891853  | 0.021045558  |
| 2                             |                   |                |              |
| ##                            | cardiacarrhythmia | cerebrovascula | chroniclr    |
| ## num_deaths                 | -0.069840050      | -0.060719123   | -0.058408066 |
| ## age45-54                   | -0.029507102      | -0.035876204   | -0.002919423 |
| ## adultrespdistSyn           | -0.043875991      | -0.039840954   | -0.043443530 |
| ## allothcondsandcauses       | -0.050851001      | -0.046174509   | -0.050349791 |
| ## alzheimer                  | -0.056553964      | -0.051353002   | -0.055996543 |
| ## cardiacArrest              | -0.038399827      | -0.034868402   | -0.038021341 |
| ## cardiacarrhythmia          | 1.000000000       | -0.036709288   | -0.040028687 |
| ## cerebrovascula             | -0.036709288      | 1.000000000    | -0.036347465 |
| ## chroniclr                  | -0.040028687      | -0.036347465   | 1.000000000  |
| ## covid19                    | -0.051180093      | -0.046473337   | -0.050675639 |
| ## diabetes                   | -0.043129400      | -0.039163022   | -0.042704298 |
| ## heartfailure               | -0.036709288      | -0.033333333   | -0.036347465 |
| ## hypertensive               | -0.040822333      | -0.037068124   | -0.040419970 |
| ## influenza                  | -0.046411569      | -0.042143348   | -0.045954116 |
| ## injuryPoisoning            | -0.041603199      | -0.037777178   | -0.041193140 |
| ## ischemicheart              | -0.038399827      | -0.034868402   | -0.038021341 |
| ## malignant                  | -0.037562786      | -0.034108340   | -0.037192551 |
| ## obesity                    | -0.041214317      | -0.037424059   | -0.040808091 |
| ## othercirculatorysystem     | -0.037562786      | -0.034108340   | -0.037192551 |
| ## otherrespystem             | -0.036709288      | -0.033333333   | -0.036347465 |
| ## renalfailure               | -0.041603199      | -0.037777178   | -0.041193140 |
| ## conditionrespiratoryarrest | -0.038399827      | -0.034868402   | -0.038021341 |
| ## sepsis                     | -0.041603199      | -0.037777178   | -0.041193140 |

```

## vascular          -0.052480839  -0.047654460  -0.051963564
## Northeast        -0.003581976   0.004048197  -0.001726085
## South            -0.028419156   -0.022477231  -0.025989362
## West             0.025922271    0.031886773   0.011080581
## covid19          diabetes      heartfailure   hypertens
ive
## num_deaths       0.346134375   0.02096477  -0.060615836   0.007835
873
## age45-54         0.005468677   0.05406184  -0.044175923   0.027104
516
## adultrespdistSyn
882
## allothcondsandcauses
072
## alzheimer        -0.071596360   -0.06033416  -0.051353002  -0.057106
783
## cardiacArrest    -0.048613530   -0.04096656  -0.034868402  -0.038775
187
## cardiacarrhythmia
333
## cerebrovascula   -0.046473337   -0.03916302  -0.033333333  -0.037068
124
## chroniclr        -0.050675639   -0.04270430  -0.036347465  -0.040419
970
## covid19           1.000000000  -0.05460109  -0.046473337  -0.051680
382
## diabetes          -0.054601090   1.000000000  -0.039163022  -0.043550
993
## heartfailure      -0.046473337   -0.03916302  1.000000000  -0.037068
124
## hypertensive      -0.051680382   -0.04355099  -0.037068124  1.000000
000
## influenza          -0.058756260   -0.04951383  -0.042143348  -0.046865
245
## injuryPoisoning   -0.052668945   -0.04438405  -0.037777178  -0.042009
873
## ischemicheart     -0.048613530   -0.04096656  -0.034868402  -0.038775
187
## malignant          -0.047553851   -0.04007357  -0.034108340  -0.037929
965
## obesity            -0.052176627   -0.04396918  -0.037424059  -0.041617
190
## othercirculatorysystem
965
## otherrespsystem    -0.047553851   -0.04007357  -0.034108340  -0.037929
124
## renalfailure       -0.046473337   -0.03916302  -0.033333333  -0.037068
873
## conditionrespiratoryarrest
187

```

```

## sepsis           -0.052668945 -0.04438405 -0.037777178 -0.042009
873
## vascular       -0.066439852 -0.05598878 -0.047654460 -0.052993
842
## Northeast      -0.025097654  0.02031950 -0.006409646  0.023022
953
## South          0.023184301  0.02402374 -0.013657052  0.017144
236
## West           -0.027178939 -0.03922366  0.041300963 -0.010681
623
##
## influenza      influenza injuryPoisoning ischemicheart
## num_deaths      0.13892821   -0.062785522 -0.054930444
## age45-54        0.01334356   -0.003034258 -0.034634398
## adultrespdistSyn -0.05037094  -0.045152364 -0.041675712
## allothcondsandcauses -0.05837845  -0.052330280 -0.048300941
## alzheimer       -0.06492562  -0.058199145 -0.053717914
## cardiacArrest    -0.04408414  -0.039516895 -0.036474164
## cardiacarrhythmia -0.04641157  -0.041603199 -0.038399827
## cerebrovascula   -0.04214335  -0.037777178 -0.034868402
## chroniclr        -0.04595412  -0.041193140 -0.038021341
## covid19          -0.05875626  -0.052668945 -0.048613530
## diabetes         -0.04951383  -0.044384054 -0.040966561
## heartfailure     -0.04214335  -0.037777178 -0.034868402
## hypertensive      -0.04686525  -0.042009873 -0.038775187
## influenza         1.00000000  -0.047761703 -0.044084136
## injuryPoisoning   -0.04776170  1.000000000 -0.039516895
## ischemicheart     -0.04408414  -0.039516895  1.000000000
## malignant        -0.04312319  -0.038655505 -0.035679100
## obesity           -0.04731525  -0.042413261 -0.039147515
## othercirculatorysystem -0.04312319  -0.038655505 -0.035679100
## otherrespystem    -0.04214335  -0.037777178 -0.034868402
## renalfailure      -0.04776170  -0.042813456 -0.039516895
## conditionrespiratoryarrest -0.04408414  -0.039516895 -0.036474164
## sepsis            -0.04776170  -0.042813456 -0.039516895
## vascular          -0.06024955  -0.054007530 -0.049849046
## Northeast         -0.02113395  0.009667327 -0.003969961
## South             0.05109605  -0.019761765 -0.015823715
## West              -0.01678239  0.010226743  0.002991381
##                         malignant   obesity othercirculatorysystem
m
## num_deaths        -0.062753494 -0.004495958           -0.043748050
6
## age45-54          -0.018986460  0.023085275           -0.002739577
1
## adultrespdistSyn -0.040767264 -0.044730307           -0.040767263
8
## allothcondsandcauses -0.047248076 -0.051841127           -0.047248075
6
## alzheimer         -0.052546969 -0.057655134           -0.052546969
4

```

|                               |                 |               |              |
|-------------------------------|-----------------|---------------|--------------|
| ## cardiacArrest              | -0.035679100    | -0.039147515  | -0.035679099 |
| 5                             |                 |               |              |
| ## cardiacarrhythmia          | -0.037562786    | -0.041214317  | -0.037562786 |
| 3                             |                 |               |              |
| ## cerebrovascula             | -0.034108340    | -0.037424059  | -0.034108340 |
| 0                             |                 |               |              |
| ## chroniclr                  | -0.037192551    | -0.040808091  | -0.037192550 |
| 8                             |                 |               |              |
| ## covid19                    | -0.047553851    | -0.052176627  | -0.047553850 |
| 9                             |                 |               |              |
| ## diabetes                   | -0.040073571    | -0.043969178  | -0.040073570 |
| 6                             |                 |               |              |
| ## heartfailure               | -0.034108340    | -0.037424059  | -0.034108340 |
| 0                             |                 |               |              |
| ## hypertensive               | -0.037929965    | -0.041617190  | -0.037929965 |
| 1                             |                 |               |              |
| ## influenza                  | -0.043123189    | -0.047315255  | -0.043123189 |
| 2                             |                 |               |              |
| ## injuryPoisoning            | -0.038655505    | -0.042413261  | -0.038655505 |
| 1                             |                 |               |              |
| ## ischemicheart              | -0.035679100    | -0.039147515  | -0.035679099 |
| 5                             |                 |               |              |
| ## malignant                  | 1.000000000     | -0.038294176  | -0.034901365 |
| 7                             |                 |               |              |
| ## obesity                    | -0.038294176    | 1.000000000   | -0.038294176 |
| 3                             |                 |               |              |
| ## othercirculatorysystem     | -0.034901366    | -0.038294176  | 1.000000000  |
| 0                             |                 |               |              |
| ## otherrespsystem            | -0.034108340    | -0.037424059  | -0.034108340 |
| 0                             |                 |               |              |
| ## renalfailure               | -0.038655505    | -0.042413261  | -0.038655505 |
| 1                             |                 |               |              |
| ## conditionrespiratoryarrest | -0.035679100    | -0.039147515  | -0.035679099 |
| 5                             |                 |               |              |
| ## sepsis                     | -0.038655505    | -0.042413261  | -0.038655505 |
| 1                             |                 |               |              |
| ## vascular                   | -0.048762435    | -0.053502700  | -0.048762435 |
| 4                             |                 |               |              |
| ## Northeast                  | 0.030662164     | 0.011589748   | -0.000045025 |
| 2                             |                 |               |              |
| ## South                      | -0.053607201    | 0.014374752   | 0.006822734  |
| 6                             |                 |               |              |
| ## West                       | 0.007917223     | -0.012819800  | 0.007917223  |
| 1                             |                 |               |              |
| ##                            | otherrespsystem | renalfailure  |              |
| ## num_deaths                 | -0.05736231     | -0.0271461038 |              |
| ## age45-54                   | -0.01927677     | 0.0634810050  |              |
| ## adultrespdistSyn           | -0.03984095     | -0.0451523641 |              |
| ## allothcondsandcauses       | -0.04617451     | -0.0523302796 |              |
| ## alzheimer                  | -0.05135300     | -0.0581991450 |              |

|                               |             |                            |                      |
|-------------------------------|-------------|----------------------------|----------------------|
| ## cardiacArrest              | -0.03486840 | -0.0395168952              |                      |
| ## cardiacarrhythmia          | -0.03670929 | -0.0416031994              |                      |
| ## cerebrovascula             | -0.03333333 | -0.0377771781              |                      |
| ## chroniclr                  | -0.03634746 | -0.0411931398              |                      |
| ## covid19                    | -0.04647334 | -0.0526689454              |                      |
| ## diabetes                   | -0.03916302 | -0.0443840543              |                      |
| ## heartfailure               | -0.03333333 | -0.0377771781              |                      |
| ## hypertensive               | -0.03706812 | -0.0420098735              |                      |
| ## influenza                  | -0.04214335 | -0.0477617029              |                      |
| ## injuryPoisoning            | -0.03777718 | -0.0428134557              |                      |
| ## ischemicheart              | -0.03486840 | -0.0395168952              |                      |
| ## malignant                  | -0.03410834 | -0.0386555051              |                      |
| ## obesity                    | -0.03742406 | -0.0424132608              |                      |
| ## othercirculatorysystem     | -0.03410834 | -0.0386555051              |                      |
| ## otherrespystem             | 1.00000000  | -0.0377771781              |                      |
| ## renalfailure               | -0.03777718 | 1.0000000000               |                      |
| ## conditionrespiratoryarrest | -0.03486840 | -0.0395168952              |                      |
| ## sepsis                     | -0.03777718 | -0.0428134557              |                      |
| ## vascular                   | -0.04765446 | -0.0540075305              |                      |
| ## Northeast                  | 0.01450604  | 0.0003550148               |                      |
| ## South                      | -0.03129741 | -0.0119077303              |                      |
| ## West                       | 0.03188677  | 0.0269927007               |                      |
| ##                            |             | conditionrespiratoryarrest | sepsis               |
| ular                          |             |                            | vasc                 |
| ## num_deaths                 |             | -0.068944526               | -0.01491374 -0.10194 |
| 1448                          |             | -0.010759068               | 0.01174691 -0.03683  |
| ## age45-54                   |             | -0.041675712               | -0.04515236 -0.05695 |
| 2946                          |             | -0.048300941               | -0.05233028 -0.06601 |
| ## adultrespdistSyn           |             | -0.053717914               | -0.05819915 -0.07341 |
| 7974                          |             | -0.036474164               | -0.03951690 -0.04984 |
| ## allothcondsandcauses       |             | -0.038399827               | -0.04160320 -0.05248 |
| 2638                          |             | -0.034868402               | -0.03777718 -0.04765 |
| ## alzheimer                  |             | -0.038021341               | -0.04119314 -0.05196 |
| 5987                          |             | -0.048613530               | -0.05266895 -0.06643 |
| ## cardiacArrest              |             | -0.040966561               | -0.04438405 -0.05598 |
| 9046                          |             | -0.034868402               | -0.03777718 -0.04765 |
| ## cardiacarrhythmia          |             | -0.038775187               | -0.04200987 -0.05299 |
| 0839                          |             |                            |                      |
| ## cerebrovascula             |             |                            |                      |
| 4460                          |             |                            |                      |
| ## chroniclr                  |             |                            |                      |
| 3564                          |             |                            |                      |
| ## covid19                    |             |                            |                      |
| 9852                          |             |                            |                      |
| ## diabetes                   |             |                            |                      |
| 8780                          |             |                            |                      |
| ## heartfailure               |             |                            |                      |
| 4460                          |             |                            |                      |
| ## hypertensive               |             |                            |                      |
| 3842                          |             |                            |                      |

|                               |               |               |              |
|-------------------------------|---------------|---------------|--------------|
| ## influenza                  | -0.044084136  | -0.04776170   | -0.06024     |
| 9554                          |               |               |              |
| ## injuryPoisoning            | -0.039516895  | -0.04281346   | -0.05400     |
| 7530                          |               |               |              |
| ## ischemicheart              | -0.036474164  | -0.03951690   | -0.04984     |
| 9046                          |               |               |              |
| ## malignant                  | -0.035679100  | -0.03865551   | -0.04876     |
| 2435                          |               |               |              |
| ## obesity                    | -0.039147515  | -0.04241326   | -0.05350     |
| 2700                          |               |               |              |
| ## othercirculatorysystem     | -0.035679100  | -0.03865551   | -0.04876     |
| 2435                          |               |               |              |
| ## otherrespystem             | -0.034868402  | -0.03777718   | -0.04765     |
| 4460                          |               |               |              |
| ## renalfailure               | -0.039516895  | -0.04281346   | -0.05400     |
| 7530                          |               |               |              |
| ## conditionrespiratoryarrest | 1.000000000   | -0.03951690   | -0.04984     |
| 9046                          |               |               |              |
| ## sepsis                     | -0.039516895  | 1.000000000   | -0.05400     |
| 7530                          |               |               |              |
| ## vascular                   | -0.049849046  | -0.05400753   | 1.00000      |
| 0000                          |               |               |              |
| ## Northeast                  | 0.016085694   | 0.01897964    | -0.01535     |
| 0020                          |               |               |              |
| ## South                      | -0.058111229  | 0.03521648    | -0.00488     |
| 5821                          |               |               |              |
| ## West                       | -0.006035707  | -0.02330517   | -0.00654     |
| 7276                          |               |               |              |
| ##                            |               |               |              |
|                               | Northeast     | South         | West         |
| ## num_deaths                 | 0.0105189052  | 0.0831565833  | -0.029851610 |
| ## age45-54                   | 0.0315721275  | 0.0442429431  | -0.071994896 |
| ## adultrespdistSyn           | -0.0012096303 | 0.0187037799  | -0.002903770 |
| ## allothcondsandcauses       | -0.0081722053 | 0.0189278425  | -0.025486637 |
| ## alzheimer                  | -0.0111635365 | -0.0003857318 | -0.008889185 |
| ## cardiacArrest              | 0.0060578666  | 0.0010912907  | 0.021045558  |
| ## cardiacarrhythmia          | -0.0035819760 | -0.0284191564 | 0.025922271  |
| ## cerebrovascula             | 0.0040481973  | -0.0224772308 | 0.031886773  |
| ## chroniclr                  | -0.0017260855 | -0.0259893622 | 0.011080581  |
| ## covid19                    | -0.0250976538 | 0.0231843009  | -0.027178939 |
| ## diabetes                   | 0.0203194976  | 0.0240237429  | -0.039223660 |
| ## heartfailure               | -0.0064096458 | -0.0136570516 | 0.041300963  |
| ## hypertensive               | 0.0230229533  | 0.0171442359  | -0.010681623 |
| ## influenza                  | -0.0211339525 | 0.0510960467  | -0.016782390 |
| ## injuryPoisoning            | 0.0096673273  | -0.0197617652 | 0.010226743  |
| ## ischemicheart              | -0.0039699611 | -0.0158237148 | 0.002991381  |
| ## malignant                  | 0.0306621636  | -0.0536072007 | 0.007917223  |
| ## obesity                    | 0.0115897483  | 0.0143747525  | -0.012819800 |
| ## othercirculatorysystem     | -0.0000450252 | 0.0068227346  | 0.007917223  |
| ## otherrespystem             | 0.0145060404  | -0.0312974099 | 0.031886773  |
| ## renalfailure               | 0.0003550148  | -0.0119077303 | 0.026992701  |

```

## conditionrespiratoryarrest  0.0160856943 -0.0581112285 -0.006035707
## sepsis                      0.0189796398  0.0352164789 -0.023305173
## vascular                    -0.0153500197 -0.0048858210 -0.006547276
## Northeast                   1.0000000000 -0.3467410014 -0.295417835
## South                        -0.3467410014  1.0000000000 -0.420859637
## West                         -0.2954178349 -0.4208596369  1.000000000

```

## QA model

```

model <- lm(num_deaths ~ ., data = covidcleansed)
summary(model)

##
## Call:
## lm(formula = num_deaths ~ ., data = covidcleansed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -147.71  -22.87   -5.79    8.37 1025.98 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                31.568     9.397   3.359 0.000804 ***
## `age45-54`                  27.471     3.850   7.136 1.57e-12 ***
## adultrespdistSyn          -22.319    12.227  -1.825 0.068178 .  
## allothcondsandcauses      21.036    11.406   1.844 0.065355 .  
## alzheimer                 -56.099   10.921  -5.137 3.21e-07 ***
## cardiacArrest              -27.852   13.151  -2.118 0.034375 *  
## cardiacarrhythmia         -51.020   12.785  -3.990 6.95e-05 *** 
## cerebrovascula            -49.770   13.516  -3.682 0.000240 *** 
## chroniclr                  -48.238   12.850  -3.754 0.000182 *** 
## covid19                     83.819   11.373   7.370 2.98e-13 *** 
## diabetes                   -21.640   12.347  -1.753 0.079896 .  
## heartfailure               -49.409   13.518  -3.655 0.000267 *** 
## hypertensive                -25.368   12.709  -1.996 0.046129 *  
## influenza                   21.204   11.890   1.783 0.074749 .  
## injuryPoisoning            -49.605   12.581  -3.943 8.47e-05 *** 
## ischemicheart              -45.561   13.158  -3.462 0.000552 *** 
## malignant                  -50.217   13.346  -3.763 0.000175 *** 
## obesity                     -29.447   12.639  -2.330 0.019960 *  
## othercirculatorysystem     -44.728   13.316  -3.359 0.000804 *** 
## otherrespystem              -49.512   13.516  -3.663 0.000259 *** 
## renalfailure                -40.641   12.589  -3.228 0.001275 ** 
## conditionrespiratoryarrest -51.853   13.182  -3.934 8.79e-05 *** 
## sepsis                      -33.406   12.578  -2.656 0.008003 ** 
## vascular                    -54.428   11.261  -4.833 1.50e-06 *** 
## Northeast                   14.850    6.008   2.471 0.013580 *  
## South                        16.162    5.358   3.016 0.002607 ** 
## West                         12.256    5.603   2.188 0.028876 *  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

## 
## Residual standard error: 70.33 on 1337 degrees of freedom
## Multiple R-squared:  0.249, Adjusted R-squared:  0.2344
## F-statistic: 17.05 on 26 and 1337 DF,  p-value: < 2.2e-16

#vif(model)

```

table1

Shows descriptive statistics about your data

```

{r} # CreateTableOne(vars = colnames(covid_clean) #
, data = covid_clean) #

features selection using all subsets selection

building the full model with all the explanatory variables

Adjusted R-squared: 0.234

fullmodel <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West, data = covidcleansed)

summary(fullmodel)

## 
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##   influenza + injuryPoisoning + ischemicheart + malignant +
##   obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##   conditionrespiratoryarrest + sepsis + vascular + Northeast +
##   South + West, data = covidcleansed)
## 

## Residuals:
##      Min    1Q Median    3Q   Max
## -147.71 -22.87  -5.79   8.37 1025.98
## 

## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 31.568     9.397   3.359  0.000804 ***
## `age45-54`                  27.471     3.850   7.136 1.57e-12 ***
## adultrespdistSyn            -22.319    12.227  -1.825  0.068178 .  
## allothcondsandcauses        21.036    11.406   1.844  0.065355 .  
## alzheimer                   -56.099    10.921  -5.137 3.21e-07 ***
```

```

## cardiacArrest          -27.852   13.151  -2.118  0.034375 *
## cardiacarrhythmia     -51.020   12.785  -3.990  6.95e-05 ***
## cerebrovascula         -49.770   13.516  -3.682  0.000240 ***
## chroniclr              -48.238   12.850  -3.754  0.000182 ***
## covid19                 83.819   11.373   7.370  2.98e-13 ***
## diabetes                -21.640   12.347  -1.753  0.079896 .
## heartfailure            -49.409   13.518  -3.655  0.000267 ***
## hypertensive             -25.368   12.709  -1.996  0.046129 *
## influenza                21.204   11.890   1.783  0.074749 .
## injuryPoisoning          -49.605   12.581  -3.943  8.47e-05 ***
## ischemicheart            -45.561   13.158  -3.462  0.000552 ***
## malignant                -50.217   13.346  -3.763  0.000175 ***
## obesity                  -29.447   12.639  -2.330  0.019960 *
## othercirculatorysystem    -44.728   13.316  -3.359  0.000804 ***
## otherrespsystem           -49.512   13.516  -3.663  0.000259 ***
## renalfailure              -40.641   12.589  -3.228  0.001275 **
## conditionrespiratoryarrest -51.853   13.182  -3.934  8.79e-05 ***
## sepsis                   -33.406   12.578  -2.656  0.008003 **
## vascular                 -54.428   11.261  -4.833  1.50e-06 ***
## Northeast                  14.850    6.008   2.471  0.013580 *
## South                      16.162    5.358   3.016  0.002607 **
## West                       12.256    5.603   2.188  0.028876 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.33 on 1337 degrees of freedom
## Multiple R-squared:  0.249, Adjusted R-squared:  0.2344
## F-statistic: 17.05 on 26 and 1337 DF, p-value: < 2.2e-16

```

## Prune diabetes

Adjusted R-squared: 0.233

```

model.covid19_deaths <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allot
hcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West, data = covidcleansed)
summary(model.covid19_deaths)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allot
## hcondsandcauses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + heartfailure + hypertensive + influenza +
##   injuryPoisoning + ischemicheart + malignant + obesity + othercirculatorysystem +

```

```

##      otherrespystem + renalfailure + conditionrespiratoryarrest +
##      sepsis + vascular + Northeast + South + West, data = covidcleansed)
##
## Residuals:
##      Min      1Q   Median      3Q     Max
## -147.37  -22.97   -6.04    8.23 1026.03
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                21.761    7.556   2.880 0.004040 **
## `age45-54`                  27.277    3.851   7.083 2.28e-12 ***
## adultrespdistSyn          -12.418   10.852  -1.144 0.252736
## allothcondsandcauses       30.933    9.917   3.119 0.001852 **
## alzheimer                  -46.206   9.357  -4.938 8.88e-07 ***
## cardiacArrest              -17.947   11.884  -1.510 0.131241
## cardiacarrhythmia          -41.142   11.485  -3.582 0.000353 ***
## cerebrovascula             -39.896   12.295  -3.245 0.001204 **
## chroniclr                  -38.342   11.552  -3.319 0.000927 ***
## covid19                     93.714    9.880   9.485 < 2e-16 ***
## heartfailure               -39.549   12.301  -3.215 0.001335 **
## hypertensive                -15.446   11.387  -1.356 0.175197
## influenza                   31.098   10.472   2.970 0.003035 **
## injuryPoisoning            -39.704   11.250  -3.529 0.000431 ***
## ischemicheart              -35.681   11.899  -2.999 0.002762 **
## malignant                  -40.309   12.099  -3.332 0.000887 ***
## obesity                     -19.531   11.310  -1.727 0.084431 .
## othercirculatorysystem     -34.835   12.070  -2.886 0.003962 **
## otherrespystem              -39.623   12.291  -3.224 0.001296 **
## renalfailure                -30.720   11.253  -2.730 0.006418 **
## conditionrespiratoryarrest -41.944   11.917  -3.520 0.000446 ***
## sepsis                      -23.492   11.243  -2.089 0.036852 *
## vascular                    -44.547    9.756  -4.566 5.42e-06 ***
## Northeast                   14.618    6.012   2.432 0.015160 *
## South                       16.213    5.362   3.024 0.002545 **
## West                        12.390    5.606   2.210 0.027281 *
##
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.38 on 1338 degrees of freedom
## Multiple R-squared:  0.2473, Adjusted R-squared:  0.2333
## F-statistic: 17.59 on 25 and 1338 DF,  p-value: < 2.2e-16

```

Prune adultrespdistSyn

Adjusted R-squared: 0.233

```

##covid_clean$ageregion <- covid_clean$age_group:covid_clean$region
fullmodel <- lm( num_deaths ~ `age45-54` + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + heartfailure + hypertensive + influenza +

```

```

injuryPoisoning + ischemicheart + malignant + obesity + othercirculatory
system +
  otherrespsystem + renalfailure + conditionrespiratoryarrest +
  sepsis + vascular + Northeast + South + West, data = covidcleansed)
summary(fullmodel)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + allothcondsandcauses +
##     alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##     chroniclr + covid19 + heartfailure + hypertensive + influenza +
##     injuryPoisoning + ischemicheart + malignant + obesity + othercirculatory
##     system +
##     otherrespsystem + renalfailure + conditionrespiratoryarrest +
##     sepsis + vascular + Northeast + South + West, data = covidcleansed)
##
## Residuals:
##    Min      1Q   Median      3Q      Max
## -147.38  -22.97   -5.42    8.24 1025.98
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.768     6.702   2.651 0.008121 **
## `age45-54`                   27.348     3.851   7.101 2.00e-12 ***
## allothcondsandcauses       34.925     9.284   3.762 0.000176 ***
## alzheimer                  -42.211     8.682  -4.862 1.30e-06 ***
## cardiacArrest                -13.950    11.361  -1.228 0.219690
## cardiacarrhythmia           -37.137    10.940  -3.395 0.000707 ***
## cerebrovascula              -35.887    11.786  -3.045 0.002374 **
## chroniclr                   -34.344    11.012  -3.119 0.001855 **
## covid19                      97.703     9.246  10.567 < 2e-16 ***
## heartfailure                 -35.537    11.792  -3.014 0.002630 **
## hypertensive                  -11.453    10.841  -1.057 0.290922
## influenza                     35.088     9.876   3.553 0.000394 ***
## injuryPoisoning              -35.705    10.695  -3.339 0.000865 ***
## ischemicheart                 -31.679    11.375  -2.785 0.005428 **
## malignant                     -36.306    11.584  -3.134 0.001760 **
## obesity                       -15.540    10.760  -1.444 0.148923
## othercirculatorysystem       -30.837    11.554  -2.669 0.007701 **
## otherrespsystem               -35.615    11.783  -3.023 0.002554 **
## renalfailure                  -26.730    10.700  -2.498 0.012608 *
## conditionrespiratoryarrest   -37.947    11.395  -3.330 0.000891 ***
## sepsis                        -19.500    10.689  -1.824 0.068335 .
## vascular                      -40.548     9.109  -4.451 9.24e-06 ***
## Northeast                      14.557     6.012   2.421 0.015598 *
## South                          16.204     5.363   3.021 0.002563 **
## West                           12.301     5.607   2.194 0.028401 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

```

## Residual standard error: 70.39 on 1339 degrees of freedom
## Multiple R-squared:  0.2466, Adjusted R-squared:  0.2331
## F-statistic: 18.26 on 24 and 1339 DF,  p-value: < 2.2e-16

```

Prune hypertensive

Adjusted R-squared: 0.233

```

fullmodel <- lm(num_deaths ~ `age45-54` + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + heartfailure + influenza +
  injuryPoisoning + ischemicheart + malignant + obesity + othercirculatorysystem +
  otherrespsystem + renalfailure + conditionrespiratoryarrest +
  sepsis + vascular + Northeast + South + West, data = covidcleansed)
summary(fullmodel)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + allothcondsandcauses +
##     alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##     chroniclr + covid19 + heartfailure + influenza + injuryPoisoning +
##     ischemicheart + malignant + obesity + othercirculatorysystem +
##     otherrespsystem + renalfailure + conditionrespiratoryarrest +
##     sepsis + vascular + Northeast + South + West, data = covidcleansed)
##
## Residuals:
##      Min      1Q      Median      3Q      Max 
## -147.26   -23.27    -5.53     8.32  1025.97 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 15.346    6.298   2.436  0.01496 *  
## `age45-54`                  27.336    3.851   7.098 2.05e-12 *** 
## allothcondsandcauses       37.424    8.978   4.168 3.27e-05 *** 
## alzheimer                   -39.713   8.354  -4.754 2.21e-06 *** 
## cardiacArrest                -11.440  11.110  -1.030  0.30331  
## cardiacarrhythmia            -34.635  10.681  -3.243  0.00121 **  
## cerebrovascula               -33.381  11.546  -2.891  0.00390 **  
## chroniclr                    -31.844  10.755  -2.961  0.00312 **  
## covid19                      100.197   8.940  11.208 < 2e-16 *** 
## heartfailure                 -33.032  11.552  -2.859  0.00431 **  
## influenza                     37.588   9.589   3.920 9.31e-05 *** 
## injuryPoisoning              -33.201  10.429  -3.183  0.00149 **  
## ischemicheart                -29.181  11.127  -2.623  0.00883 **  
## malignant                     -33.798  11.338  -2.981  0.00293 **  
## obesity                       -13.033  10.496  -1.242  0.21455  
## othercirculatorysystem        -28.332  11.309  -2.505  0.01235 *  
## otherrespsystem                -33.106  11.542  -2.868  0.00419 **  
## renalfailure                  -24.222  10.434  -2.321  0.02041 *  
## conditionrespiratoryarrest   -35.448  11.147  -3.180  0.00151 ** 

```

```

## sepsis           -16.991    10.422   -1.630  0.10329
## vascular        -38.052     8.798   -4.325  1.64e-05 ***
## Northeast       14.385     6.010    2.393  0.01683 *
## South            16.149     5.363    3.011  0.00265 **
## West             12.216     5.606    2.179  0.02951 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.39 on 1340 degrees of freedom
## Multiple R-squared:  0.246, Adjusted R-squared:  0.233
## F-statistic:    19 on 23 and 1340 DF, p-value: < 2.2e-16

##Prune cardiacArrest ## Adjusted R-squared: 0.233

fullmodel <- lm(num_deaths ~ `age45-54` + allothcondsandcauses +
  alzheimer + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + heartfailure + influenza +
  injuryPoisoning + ischemicheart + malignant + obesity + othercirculatorysystem +
  otherrespsystem + renalfailure + conditionrespiratoryarrest +
  sepsis + vascular + Northeast + South + West, data = covidcleansed)
summary(fullmodel)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + allothcondsandcauses +
##     alzheimer + cardiacarrhythmia + cerebrovascula + chroniclr +
##     covid19 + heartfailure + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + South + West, data = covidcleansed)
##
## Residuals:
##      Min      1Q      Median      3Q      Max 
## -147.23  -23.85   -5.74    8.34 1025.93 
##
## Coefficients:
## (Intercept)          13.542    6.050   2.238  0.02536 * 
## `age45-54`           27.334    3.851   7.097 2.06e-12 ***
## allothcondsandcauses 39.279    8.796   4.466 8.65e-06 ***
## alzheimer            -37.853   8.157   -4.641 3.81e-06 *** 
## cardiacarrhythmia    -32.764   10.526  -3.113  0.00189 ** 
## cerebrovascula       -31.505   11.401  -2.763  0.00580 ** 
## chroniclr            -29.977   10.602  -2.828  0.00476 ** 
## covid19               102.050   8.757   11.653 < 2e-16 ***
## heartfailure         -31.154   11.407  -2.731  0.00640 ** 
## influenza             39.444    9.418   4.188  3.00e-05 *** 
## injuryPoisoning      -31.332   10.271  -3.051  0.00233 ** 
## ischemicheart        -27.317   10.979  -2.488  0.01296 * 

```

```

## malignant             -31.928   11.192  -2.853  0.00440  **
## obesity               -11.172   10.340  -1.081  0.28009
## othercirculatorysystem -26.466   11.163  -2.371  0.01789  *
## otherrespsystem        -31.228   11.397  -2.740  0.00622  **
## renalfailure           -22.350   10.275  -2.175  0.02979  *
## conditionrespiratoryarrest -33.585   10.999  -3.053  0.00231  **
## sepsis                 -15.132   10.265  -1.474  0.14069
## vascular               -36.193    8.611  -4.203  2.81e-05 ***
## Northeast              14.309    6.010   2.381  0.01741  *
## South                  16.144    5.363   3.010  0.00266  **
## West                   12.057    5.604   2.151  0.03162  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.39 on 1341 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.233
## F-statistic: 19.82 on 22 and 1341 DF,  p-value: < 2.2e-16

```

## Prune obesity

Adjusted R-squared: 0.233

```

fullmodel <- lm(num_deaths ~ `age45-54` + allothcondsandcauses +
  alzheimer + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + heartfailure + influenza +
  injuryPoisoning + ischemicheart + malignant + othercirculatorysystem +
  otherrespsystem + renalfailure + conditionrespiratoryarrest +
  sepsis + vascular + Northeast + South + West, data = covidcleansed)
summary(fullmodel)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + allothcondsandcauses +
##     alzheimer + cardiacarrhythmia + cerebrovascula + chroniclr +
##     covid19 + heartfailure + influenza + injuryPoisoning + ischemicheart +
##     malignant + othercirculatorysystem + otherrespsystem + renalfailure +
##     conditionrespiratoryarrest + sepsis + vascular + Northeast +
##     South + West, data = covidcleansed)
##
## Residuals:
##      Min      1Q      Median      3Q      Max 
## -147.20  -23.86   -6.28    8.38  1025.91 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                11.773    5.825   2.021  0.04346 *  
## `age45-54`                  27.340    3.852   7.098 2.04e-12 ***
## allothcondsandcauses       41.036    8.645   4.747 2.29e-06 ***
## alzheimer                  -36.096    7.994  -4.516 6.87e-06 *** 
## cardiacarrhythmia          -31.007   10.400  -2.981  0.00292 ** 
## cerebrovascula              -29.748   11.286  -2.636  0.00849 ** 

```

```

## chroniclr           -28.220   10.477  -2.694  0.00716  **
## covid19              103.806   8.606   12.063 < 2e-16 ***
## heartfailure        -29.399   11.292  -2.604  0.00933  **
## influenza             41.197   9.278   4.440  9.72e-06 ***
## injuryPoisoning      -29.575   10.142  -2.916  0.00360  **
## ischemicheart         -25.560   10.858  -2.354  0.01872  *
## malignant             -30.167   11.074  -2.724  0.00653  **
## othercirculatorysystem -24.710   11.045  -2.237  0.02543  *
## otherrespsystem        -29.471   11.281  -2.612  0.00909  **
## renalfailure           -20.595   10.146  -2.030  0.04258  *
## conditionrespiratoryarrest -31.823   10.879  -2.925  0.00350  **
## sepsis                 -13.375   10.136  -1.320  0.18721
## vascular               -34.436    8.456  -4.072  4.93e-05 ***
## Northeast                14.283   6.010   2.377  0.01762  *
## South                   16.169   5.363   3.015  0.00262  **
## West                     12.083   5.604   2.156  0.03126  *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.4 on 1342 degrees of freedom
## Multiple R-squared:  0.2447, Adjusted R-squared:  0.2329
## F-statistic:  20.7 on 21 and 1342 DF,  p-value: < 2.2e-16

```

##Prune sepsis ##

```

fullmodel <- lm(num_deaths ~ `age45-54` + allothcondsandcauses +
  alzheimer + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + heartfailure + influenza +
  injuryPoisoning + ischemicheart + malignant + othercirculatorysystem +
  otherrespsystem + renalfailure + conditionrespiratoryarrest +
  vascular + Northeast + South + West, data = covidcleansed)
summary(fullmodel)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + allothcondsandcauses +
##     alzheimer + cardiacarrhythmia + cerebrovascula + chroniclr +
##     covid19 + heartfailure + influenza + injuryPoisoning + ischemicheart +
##     malignant + othercirculatorysystem + otherrespsystem + renalfailure +
##     conditionrespiratoryarrest + vascular + Northeast + South +
##     West, data = covidcleansed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -147.16  -23.42   -7.14    8.49 1025.94
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.977     5.665   1.761  0.07843 .
## `age45-54`                  27.404     3.852   7.114 1.83e-12 ***

```

```

## allothcondsandcauses      42.879    8.534   5.025 5.72e-07 ***
## alzheimer                 -34.257    7.874  -4.351 1.46e-05 ***
## cardiacarrhythmia        -29.170   10.309  -2.830  0.00473 **
## cerebrovascula            -27.906   11.202  -2.491  0.01285 *
## chroniclr                  -26.387   10.387  -2.540  0.01119 *
## covid19                     105.644   8.494  12.437 < 2e-16 ***
## heartfailure                -27.556   11.208  -2.459  0.01407 *
## influenza                   43.044    9.174   4.692 2.98e-06 ***
## injuryPoisoning             -27.737   10.048  -2.760  0.00585 **
## ischemicheart                -23.720   10.771  -2.202  0.02783 *
## malignant                    -28.331   10.989  -2.578  0.01004 *
## othercirculatorysystem     -22.867   10.959  -2.087  0.03711 *
## otherrespsystem              -27.632   11.198  -2.468  0.01373 *
## renalfailure                  -18.767   10.054  -1.867  0.06218 .
## conditionrespiratoryarrest -29.995   10.793  -2.779  0.00553 **
## vascular                      -32.596   8.343  -3.907 9.81e-05 ***
## Northeast                     14.132    6.011   2.351  0.01886 *
## South                          16.035    5.364   2.990  0.00284 **
## West                           12.065    5.606   2.152  0.03156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.42 on 1343 degrees of freedom
## Multiple R-squared:  0.2437, Adjusted R-squared:  0.2325
## F-statistic: 21.64 on 20 and 1343 DF,  p-value: < 2.2e-16

```

## Model selection

building a full model with everything using the backward selection

```

library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:gtsummary':
##
##     select

## The following object is masked from 'package:dplyr':
##
##     select

#model_full <- lm(num_deaths ~ age_group + condition + condition_group + region, data = covid_clean)
model_full <- lm(num_deaths ~ ., data = covidcleansed)
step <- stepAIC(model_full, direction = "backward")

## Start: AIC=11629.34
## num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
##     alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##     chroniclr + covid19 + diabetes + heartfailure + hypertensive +

```

```

## influenza + injuryPoisoning + ischemicheart + malignant +
## obesity + othercirculatorysystem + otherrespsystem + renalfailure +
## conditionrespiratoryarrest + sepsis + vascular + Northeast +
## South + West
##
##                                     Df Sum of Sq      RSS     AIC
## <none>                               6612665 11629
## - diabetes                           1    15192 6627858 11630
## - influenza                          1    15730 6628396 11631
## - adultrespdistSyn                 1    16478 6629144 11631
## - allothcondsandcauses             1    16824 6629489 11631
## - hypertensive                      1    19706 6632371 11631
## - cardiacArrest                     1    22183 6634848 11632
## - West                               1    23668 6636333 11632
## - obesity                            1    26848 6639514 11633
## - Northeast                          1    30210 6642876 11634
## - sepsis                             1    34888 6647553 11634
## - South                              1    44999 6657665 11637
## - renalfailure                      1    51547 6664212 11638
## - othercirculatorysystem            1    55804 6668470 11639
## - ischemicheart                     1    59294 6671960 11640
## - heartfailure                      1    66076 6678741 11641
## - otherrespsystem                   1    66370 6679035 11641
## - cerebrovascula                    1    67068 6679733 11641
## - chroniclr                          1    69698 6682363 11642
## - malignant                          1    70021 6682687 11642
## - conditionrespiratoryarrest       1    76533 6689199 11643
## - injuryPoisoning                  1    76887 6689552 11643
## - cardiacarrhythmia                1    78757 6691422 11644
## - vascular                           1    115542 6728207 11651
## - alzheimer                         1    130494 6743159 11654
## - `age45-54`                        1    251833 6864498 11678
## - covid19                            1    268642 6881307 11682

step$anova ## display results

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
##         alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##         chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##         influenza + injuryPoisoning + ischemicheart + malignant +
##         obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##         conditionrespiratoryarrest + sepsis + vascular + Northeast +
##         South + West
##
## Final Model:
## num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +

```

```

##      alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##      chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##      influenza + injuryPoisoning + ischemicheart + malignant +
##      obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##      conditionrespiratoryarrest + sepsis + vascular + Northeast +
##      South + West
##
##
##      Step Df Deviance Resid. Df Resid. Dev      AIC
## 1           1337    6612665 11629.34

summary(step)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##      alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##      chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##      influenza + injuryPoisoning + ischemicheart + malignant +
##      obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##      conditionrespiratoryarrest + sepsis + vascular + Northeast +
##      South + West, data = covidcleansed)
##
## Residuals:
##      Min      1Q      Median      3Q      Max
## -147.71   -22.87    -5.79     8.37 1025.98
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 31.568    9.397   3.359 0.000804 ***
## `age45-54`                  27.471    3.850   7.136 1.57e-12 ***
## adultrespdistSyn            -22.319   12.227  -1.825 0.068178 .  
## allothcondsandcauses       21.036   11.406   1.844 0.065355 .  
## alzheimer                  -56.099   10.921  -5.137 3.21e-07 ***
## cardiacArrest                -27.852   13.151  -2.118 0.034375 *  
## cardiacarrhythmia            -51.020   12.785  -3.990 6.95e-05 *** 
## cerebrovascula              -49.770   13.516  -3.682 0.000240 *** 
## chroniclr                   -48.238   12.850  -3.754 0.000182 *** 
## covid19                      83.819   11.373   7.370 2.98e-13 ***
## diabetes                     -21.640   12.347  -1.753 0.079896 .  
## heartfailure                 -49.409   13.518  -3.655 0.000267 *** 
## hypertensive                  -25.368   12.709  -1.996 0.046129 *  
## influenza                     21.204   11.890   1.783 0.074749 .  
## injuryPoisoning              -49.605   12.581  -3.943 8.47e-05 *** 
## ischemicheart                -45.561   13.158  -3.462 0.000552 *** 
## malignant                    -50.217   13.346  -3.763 0.000175 *** 
## obesity                      -29.447   12.639  -2.330 0.019960 *  
## othercirculatorysystem       -44.728   13.316  -3.359 0.000804 *** 
## otherrespsystem               -49.512   13.516  -3.663 0.000259 *** 

```

```

## renalfailure           -40.641   12.589  -3.228 0.001275 **
## conditionrespiratoryarrest -51.853   13.182  -3.934 8.79e-05 ***
## sepsis                 -33.406   12.578  -2.656 0.008003 **
## vascular                -54.428   11.261  -4.833 1.50e-06 ***
## Northeast               14.850    6.008   2.471 0.013580 *
## South                   16.162    5.358   3.016 0.002607 **
## West                    12.256    5.603   2.188 0.028876 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.33 on 1337 degrees of freedom
## Multiple R-squared:  0.249, Adjusted R-squared:  0.2344
## F-statistic: 17.05 on 26 and 1337 DF, p-value: < 2.2e-16

```

building a full model with everything using foward selection

```

model_full <- lm(num_deaths ~ ., data = covidcleansed)
model_empty <- lm(num_deaths ~ 1, data = covidcleansed)
summary(model_empty)

##
## Call:
## lm(formula = num_deaths ~ 1, data = covidcleansed)
##
## Residuals:
##     Min      1Q      Median      3Q      Max 
## -31.38  -31.38   -21.38   -4.38 1153.62 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 31.380     2.176   14.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.38 on 1363 degrees of freedom

## pass the empty model to stepAIC()
## consider every variable all the way down to no variable
## that is why we define the scope
step <- stepAIC(model_empty, direction = "forward", scope = list(upper=model_full, lower=model_empty))

## Start: AIC=11968
## num_deaths ~ 1
##
##                                     Df Sum of Sq    RSS    AIC
## + covid19                           1  1054995 7750643 11796
## + `age45-54`                         1   315224 8490413 11920
## + allothcondsandcauses              1   186849 8618789 11941
## + influenza                          1   169958 8635679 11943
## + alzheimer                         1   106264 8699373 11953

```

```

## + vascular          1  91509 8714129 11956
## + South             1  60891 8744746 11960
## + cardiacarrhythmia 1  42951 8762687 11963
## + conditionrespiratoryarrest 1  41856 8763781 11964
## + injuryPoisoning   1  34712 8770926 11965
## + malignant          1  34677 8770961 11965
## + cerebrovascula     1  32465 8773173 11965
## + heartfailure        1  32354 8773283 11965
## + chroniclr           1  30040 8775597 11965
## + otherrespsystem     1  28974 8776663 11966
## + ischemicheart       1  26570 8779068 11966
## + othercirculatorysystem 1  16853 8788784 11967
## <none>                  8805638 11968
## + West                1  7847 8797791 11969
## + renalfailure         1  6489 8799149 11969
## + diabetes              1  3870 8801767 11969
## + sepsis                1  1959 8803679 11970
## + adultrespdistSyn      1  1329 8804309 11970
## + Northeast             1  974 8804663 11970
## + hypertensive            1  541 8805097 11970
## + obesity                 1  178 8805460 11970
## + cardiacArrest          1  0 8805637 11970
##
## Step: AIC=11795.93
## num_deaths ~ covid19
##
##                                     Df Sum of Sq    RSS    AIC
## + `age45-54`                   1  308958 7441685 11742
## + allothcondsandcauses        1  249419 7501224 11753
## + influenza                     1  224134 7526509 11758
## + alzheimer                    1  64056 7686587 11787
## + vascular                      1  55122 7695521 11788
## + South                         1  49733 7700910 11789
## + cardiacarrhythmia            1  23988 7726655 11794
## + conditionrespiratoryarrest   1  23975 7726668 11794
## + malignant                     1  18914 7731729 11795
## + cerebrovascula               1  17580 7733063 11795
## + injuryPoisoning              1  17529 7733114 11795
## + heartfailure                  1  17499 7733144 11795
## + otherrespsystem               1  15035 7735608 11795
## + chroniclr                     1  14745 7735898 11795
## + diabetes                      1  14035 7736608 11796
## + ischemicheart                 1  12815 7737828 11796
## <none>                          7750643 11796
## + adultrespdistSyn             1  8770 7741873 11796
## + othercirculatorysystem        1  6572 7744071 11797
## + hypertensive                  1  5843 7744800 11797
## + West                          1  3683 7746960 11797
## + Northeast                     1  3250 7747393 11797
## + cardiacArrest                 1  2569 7748074 11798

```

```

## + obesity           1   1625 7749018 11798
## + renalfailure     1   702 7749941 11798
## + sepsis            1   97 7750546 11798
##
## Step: AIC=11742.44
## num_deaths ~ covid19 + `age45-54`
##
##                                     Df Sum of Sq    RSS    AIC
## + allothcondsandcauses  1   251293 7190392 11698
## + influenza             1   217028 7224657 11704
## + alzheimer             1   61448 7380237 11733
## + vascular              1   46056 7395629 11736
## + South                 1   39471 7402214 11737
## + conditionrespiratoryarrest 1   22203 7419482 11740
## + cardiacarrhythmia    1   19230 7422455 11741
## + injuryPoisoning      1   17127 7424558 11741
## + malignant             1   16162 7425523 11742
## + chroniclr             1   14390 7427295 11742
## + cerebrovascula        1   12733 7428953 11742
## + otherrespystem        1   12556 7429129 11742
## + heartfailure          1   11652 7430033 11742
## <none>                  7441685 11742
## + ischemicheart         1   8861 7432824 11743
## + diabetes               1   7804 7433881 11743
## + adultrespdistSyn     1   7644 7434042 11743
## + othercirculatorysystem 1   6350 7435335 11743
## + renalfailure          1   3858 7437827 11744
## + hypertensive           1   3747 7437938 11744
## + Northeast              1   1552 7440133 11744
## + cardiacArrest          1   1507 7440178 11744
## + obesity                1   746 7440940 11744
## + West                   1   432 7441253 11744
## + sepsis                 1   10 7441675 11744
##
## Step: AIC=11697.59
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses
##
##                                     Df Sum of Sq    RSS    AIC
## + influenza             1   248095 6942297 11652
## + alzheimer             1   44240 7146152 11691
## + South                 1   35483 7154910 11693
## + vascular              1   32251 7158141 11694
## + conditionrespiratoryarrest 1   15200 7175192 11697
## + diabetes               1   13818 7176574 11697
## + adultrespdistSyn     1   13731 7176661 11697
## + cardiacarrhythmia    1   12441 7177951 11697
## + injuryPoisoning      1   10608 7179784 11698
## <none>                  7190392 11698
## + malignant             1   10388 7180004 11698
## + chroniclr             1   8672 7181720 11698

```

```

## + hypertensive          1    7888 7182504 11698
## + cerebrovascula       1    7770 7182622 11698
## + otherrespsystem      1    7638 7182754 11698
## + heartfailure         1    6926 7183466 11698
## + ischemicheart        1    4663 7185729 11699
## + cardiacArrest        1    4192 7186200 11699
## + obesity               1    3039 7187353 11699
## + othercirculatorysystem 1    2965 7187427 11699
## + Northeast              1    1960 7188432 11699
## + renalfailure          1    1169 7189223 11699
## + sepsis                1     975 7189417 11699
## + West                  1      48 7190344 11700
##
## Step: AIC=11651.69
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza
##
##                                     Df Sum of Sq   RSS   AIC
## + alzheimer                 1  30188.3 6912108 11648
## + South                      1  26241.7 6916055 11648
## + diabetes                   1  21448.1 6920848 11650
## + adultrespdistSyn          1  21387.4 6920909 11650
## + vascular                   1  21235.3 6921061 11650
## + hypertensive               1  13435.3 6928861 11651
## <none>                      6942297 11652
## + conditionrespiratoryarrest 1  9671.7 6932625 11652
## + cardiacArrest              1  8126.1 6934171 11652
## + cardiacarrhythmia          1  7289.0 6935008 11652
## + obesity                     1  6789.8 6935507 11652
## + malignant                  1  6014.5 6936282 11652
## + injuryPoisoning            1  5758.3 6936538 11653
## + chroniclr                  1  4492.2 6937804 11653
## + cerebrovascula             1  4150.0 6938147 11653
## + otherrespsystem            1  4041.2 6938255 11653
## + heartfailure               1  3545.2 6938751 11653
## + sepsis                     1  3437.9 6938859 11653
## + Northeast                  1  3149.1 6939148 11653
## + ischemicheart              1  1885.3 6940411 11653
## + othercirculatorysystem     1   895.2 6941401 11654
## + renalfailure               1    41.8 6942255 11654
## + West                        1     7.0 6942290 11654
##
## Step: AIC=11647.75
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##     alzheimer
##
##                                     Df Sum of Sq   RSS   AIC
## + South                      1  26654.5 6885454 11644
## + vascular                   1  26272.4 6885836 11645
## + diabetes                   1  17971.3 6894137 11646
## + adultrespdistSyn          1  17847.1 6894261 11646

```

```

## + conditionrespiratoryarrest 1 12100.3 6900008 11647
## + hypertensive 1 10831.0 6901277 11648
## <none> 6912108 11648
## + cardiacarrhythmia 1 9539.4 6902569 11648
## + malignant 1 7906.7 6904202 11648
## + injuryPoisoning 1 7827.6 6904281 11648
## + chroniclr 1 6257.4 6905851 11648
## + cardiacArrest 1 6231.7 6905877 11648
## + cerebrovascula 1 5704.5 6906404 11649
## + otherrespsystem 1 5572.0 6906536 11649
## + heartfailure 1 4992.4 6907116 11649
## + obesity 1 4950.7 6907158 11649
## + ischemicheart 1 3025.1 6909083 11649
## + Northeast 1 2858.5 6909250 11649
## + sepsis 1 2153.6 6909955 11649
## + othercirculatorysystem 1 1694.1 6910414 11649
## + renalfailure 1 353.7 6911755 11650
## + West 1 0.0 6912108 11650
##
## Step: AIC=11644.48
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##      alzheimer + South
##
##                                     Df Sum of Sq    RSS   AIC
## + vascular 1 26488.8 6858965 11641
## + adultrespdistSyn 1 16800.6 6868653 11643
## + diabetes 1 16779.8 6868674 11643
## + Northeast 1 13807.8 6871646 11644
## + conditionrespiratoryarrest 1 10307.7 6875146 11644
## + hypertensive 1 10105.9 6875348 11644
## <none> 6885454 11644
## + cardiacarrhythmia 1 8863.1 6876591 11645
## + injuryPoisoning 1 7430.2 6878024 11645
## + malignant 1 6590.2 6878864 11645
## + cardiacArrest 1 6089.4 6879364 11645
## + chroniclr 1 5741.7 6879712 11645
## + West 1 5645.5 6879808 11645
## + cerebrovascula 1 5319.1 6880135 11645
## + otherrespsystem 1 4969.3 6880485 11646
## + heartfailure 1 4839.3 6880615 11646
## + obesity 1 4519.1 6880935 11646
## + ischemicheart 1 2864.8 6882589 11646
## + othercirculatorysystem 1 1863.3 6883591 11646
## + sepsis 1 1581.1 6883873 11646
## + renalfailure 1 300.6 6885153 11646
##
## Step: AIC=11641.22
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##      alzheimer + South + vascular
##

```

```

##                                     Df Sum of Sq    RSS   AIC
## + diabetes                         1  13887.1 6845078 11640
## + adultrespdistSyn                 1  13803.8 6845161 11640
## + Northeast                        1  13073.3 6845892 11641
## + conditionrespiratoryarrest      1  12716.1 6846249 11641
## + cardiacarrhythmia                1  11259.6 6847705 11641
## <none>                            6858965 11641
## + injuryPoisoning                  1   9669.1 6849296 11641
## + malignant                        1   8496.2 6850469 11642
## + hypertensive                      1   7967.1 6850998 11642
## + chroniclr                         1   7639.0 6851326 11642
## + cerebrovascula                   1   7018.5 6851947 11642
## + otherrespsystem                   1   6596.5 6852369 11642
## + heartfailure                      1   6474.3 6852491 11642
## + West                             1   5272.9 6853692 11642
## + cardiacArrest                     1   4538.6 6854427 11642
## + ischemicheart                    1   4202.0 6854763 11642
## + obesity                           1   3107.0 6855858 11643
## + othercirculatorysystem            1   2916.4 6856049 11643
## + renalfailure                      1     835.4 6858130 11643
## + sepsis                            1     784.6 6858181 11643
##
## Step:  AIC=11640.46
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##           alzheimer + South + vascular + diabetes
##
##                                     Df Sum of Sq    RSS   AIC
## + adultrespdistSyn                 1  15902.3 6829176 11639
## + Northeast                        1  12434.1 6832644 11640
## + conditionrespiratoryarrest      1  11250.2 6833828 11640
## <none>                            6845078 11640
## + cardiacarrhythmia                1   9811.7 6835266 11640
## + hypertensive                      1   9465.2 6835613 11641
## + injuryPoisoning                  1   8257.4 6836821 11641
## + malignant                        1   7331.9 6837746 11641
## + chroniclr                         1   6435.6 6838642 11641
## + cerebrovascula                   1   5986.1 6839092 11641
## + West                             1   5870.8 6839207 11641
## + cardiacArrest                     1   5591.4 6839487 11641
## + otherrespsystem                   1   5586.4 6839492 11641
## + heartfailure                      1   5486.3 6839592 11641
## + obesity                           1   4062.5 6841015 11642
## + ischemicheart                    1   3369.6 6841708 11642
## + othercirculatorysystem            1   2219.0 6842859 11642
## + sepsis                            1   1298.0 6843780 11642
## + renalfailure                      1     434.9 6844643 11642
##
## Step:  AIC=11639.29
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##           alzheimer + South + vascular + diabetes + adultrespdistSyn

```

```

##                                     Df Sum of Sq    RSS   AIC
## + Northeast                         1  12354.6 6816821 11639
## + hypertensive                      1  11350.7 6817825 11639
## <none>                               6829176 11639
## + conditionrespiratoryarrest        1   9630.9 6819545 11639
## + cardiacarrhythmia                1   8210.7 6820965 11640
## + cardiacArrest                     1   6944.5 6822231 11640
## + injuryPoisoning                  1   6740.5 6822435 11640
## + malignant                         1   6057.5 6823118 11640
## + West                             1   5881.4 6823294 11640
## + obesity                           1   5325.6 6823850 11640
## + chroniclr                         1   5152.6 6824023 11640
## + cerebrovascula                   1   4852.8 6824323 11640
## + otherrespsystem                  1   4494.1 6824682 11640
## + heartfailure                      1   4400.3 6824775 11640
## + ischemicheart                    1   2491.4 6826684 11641
## + sepsis                            1   2056.3 6827119 11641
## + othercirculatorysystem            1   1522.1 6827654 11641
## + renalfailure                      1     138.5 6829037 11641
##
## Step:  AIC=11638.82
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##      alzheimer + South + vascular + diabetes + adultrespdistSyn +
##      Northeast
##
##                                     Df Sum of Sq    RSS   AIC
## + West                            1  24667.4 6792154 11636
## + hypertensive                     1  10703.6 6806118 11639
## <none>                           6816821 11639
## + conditionrespiratoryarrest      1   9503.1 6807318 11639
## + cardiacarrhythmia              1   7909.7 6808911 11639
## + cardiacArrest                   1   6870.0 6809951 11639
## + injuryPoisoning                 1   6760.6 6810061 11640
## + malignant                        1   6268.3 6810553 11640
## + obesity                          1   5088.9 6811732 11640
## + chroniclr                        1   4940.1 6811881 11640
## + cerebrovascula                  1   4788.7 6812032 11640
## + otherrespsystem                  1   4539.2 6812282 11640
## + heartfailure                     1   4229.6 6812592 11640
## + ischemicheart                   1   2376.9 6814444 11640
## + sepsis                           1   1754.8 6815066 11640
## + othercirculatorysystem           1   1529.0 6815292 11640
## + renalfailure                     1     115.2 6816706 11641
##
## Step:  AIC=11635.87
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##      alzheimer + South + vascular + diabetes + adultrespdistSyn +
##      Northeast + West
##

```

```

##                                     Df Sum of Sq    RSS   AIC
## + hypertensive                   1  10479.6 6781674 11636
## <none>                           6792154 11636
## + conditionrespiratoryarrest    1   8030.8 6784123 11636
## + cardiacarrhythmia            1   7909.2 6784245 11636
## + injuryPoisoning              1   6647.8 6785506 11636
## + cardiacArrest                1   6234.0 6785920 11637
## + malignant                     1   5791.7 6786362 11637
## + obesity                       1   5219.0 6786935 11637
## + cerebrovascula               1   5187.6 6786966 11637
## + otherrespsystem               1   4942.7 6787211 11637
## + heartfailure                  1   4866.4 6787287 11637
## + chroniclr                      1   4610.5 6787543 11637
## + ischemicheart                 1   2076.7 6790077 11638
## + sepsis                         1   1741.9 6790412 11638
## + othercirculatorysystem         1   1624.7 6790529 11638
## + renalfailure                   1    193.7 6791960 11638
##
## Step:  AIC=11635.76
## num_deaths ~ covid19 + `age45-54` + allothcondsandcauses + influenza +
##      alzheimer + South + vascular + diabetes + adultrespdistSyn +
##      Northeast + West + hypertensive
##
##                                     Df Sum of Sq    RSS   AIC
## <none>                           6781674 11636
## + cardiacArrest                   1   7405.8 6774268 11636
## + conditionrespiratoryarrest     1   6923.3 6774751 11636
## + cardiacarrhythmia              1   6749.0 6774925 11636
## + obesity                        1   6388.1 6775286 11636
## + injuryPoisoning                1   5535.0 6776139 11637
## + malignant                      1   4866.5 6776808 11637
## + cerebrovascula                 1   4331.2 6777343 11637
## + otherrespsystem                 1   4101.7 6777572 11637
## + heartfailure                   1   4038.8 6777635 11637
## + chroniclr                      1   3729.1 6777945 11637
## + sepsis                          1   2449.9 6779224 11637
## + ischemicheart                  1   1523.3 6780151 11638
## + othercirculatorysystem          1   1136.1 6780538 11638
## + renalfailure                   1    40.7 6781633 11638

summary(step)

##
## Call:
## lm(formula = num_deaths ~ covid19 + `age45-54` + allothcondsandcauses +
##      influenza + alzheimer + South + vascular + diabetes + adultrespdistSyn +
##      Northeast + West + hypertensive, data = covidcleansed)
##
## Residuals:

```

```

##      Min     1Q   Median     3Q    Max
## -147.41 -23.43   -5.35    9.59 1024.27
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|) 
## (Intercept)             -9.588     4.936  -1.943 0.052277 .
## covid19                  123.669    8.200  15.082 < 2e-16 ***
## `age45-54`                28.524    3.858   7.394 2.5e-13 ***
## allothcondsandcauses     60.921    8.240   7.394 2.5e-13 ***
## influenza                 60.914    8.922   6.827 1.3e-11 ***
## alzheimer                -16.138    7.537  -2.141 0.032438 *
## South                      18.120    5.368   3.376 0.000757 ***
## vascular                  -14.395    8.024  -1.794 0.073027 .
## diabetes                   18.079    9.523   1.898 0.057844 .
## adultrespdistSyn          17.517    9.360   1.872 0.061480 .
## Northeast                  14.803    6.042   2.450 0.014418 *
## West                       12.418    5.627   2.207 0.027504 *
## hypertensive                14.430    9.987   1.445 0.148724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.85 on 1351 degrees of freedom
## Multiple R-squared:  0.2298, Adjusted R-squared:  0.223
## F-statistic: 33.6 on 12 and 1351 DF,  p-value: < 2.2e-16

model
lm_fit <- lm(num_deaths ~ condition + condition_group + age_group + region,
data = covid_clean)
summary(lm_fit)

##
## Call:
## lm(formula = num_deaths ~ condition + condition_group + age_group +
##     region, data = covid_clean)
##
## Residuals:
##      Min     1Q   Median     3Q    Max
## -147.71 -22.87   -5.79    8.37 1025.98
##
## Coefficients: (11 not defined because of singularities)
##
## Estimate
## (Intercept)           31.568
## conditionAdultRespDistSyn -22.319
## conditionAll other conditions and causes (residual) 21.036
## conditionAlzheimer   -56.099

```

```
## conditionCardiacArrest
-27.852
## conditionCardiacArrhythmia
-51.020
## conditionCerebrovascular
-49.770
## conditionChronicLR
-48.238
## conditionCOVID-19
83.819
## conditionDiabetes
-21.640
## conditionHeartFailure
-49.409
## conditionHypertensive
-25.368
## conditionInfluenza
21.204
## conditioninjuryPoisoning
-49.605
## conditionIschemicheart
-45.561
## conditionMalignant
-50.217
## conditionObesity
-29.447
## conditionOtherCirculatorySystem
-44.728
## conditionOtherRespSystem
-49.512
## conditionRenal failure
-40.641
## conditionRespiratoryArrest
-51.853
## conditionSepsis
-33.406
## conditionVascular
-54.428
## condition_groupAll other conditions and causes (residual)
NA
## condition_groupAlzheimer disease
NA
## condition_groupCirculatory diseases
NA
## condition_groupCoronavirus Disease 2019
NA
## condition_groupDiabetes
NA
## condition_groupIntentional and unintentional injury, poisoning and other a
dverse events      NA
```

```
## condition_groupMalignant neoplasms
NA
## condition_groupObesity
NA
## condition_groupRenal failure
NA
## condition_groupSepsis
NA
## condition_groupVascular and unspecified dementia
NA
## age_group45-54
27.471
## regionNortheast
14.850
## regionSouth
16.162
## regionWest
12.256
##
Std. Error
## (Intercept)
9.397
## conditionAdultRespDistSyn
12.227
## conditionAll other conditions and causes (residual)
11.406
## conditionAlzheimer
10.921
## conditionCardiacArrest
13.151
## conditionCardiacArrhythmia
12.785
## conditionCerebrovascular
13.516
## conditionChronicLR
12.850
## conditionCOVID-19
11.373
## conditionDiabetes
12.347
## conditionHeartFailure
13.518
## conditionHypertensive
12.709
## conditionInfluenza
11.890
## conditioninjuryPoisoning
12.581
## conditionIschemicheart
13.158
```

```
## conditionMalignant
13.346
## conditionObesity
12.639
## conditionOtherCirculatorySystem
13.316
## conditionOtherRespSystem
13.516
## conditionRenal failure
12.589
## conditionRespiratoryArrest
13.182
## conditionSepsis
12.578
## conditionVascular
11.261
## condition_groupAll other conditions and causes (residual)
NA
## condition_groupAlzheimer disease
NA
## condition_groupCirculatory diseases
NA
## condition_groupCoronavirus Disease 2019
NA
## condition_groupDiabetes
NA
## condition_groupIntentional and unintentional injury, poisoning and other a
dverse events NA
## condition_groupMalignant neoplasms
NA
## condition_groupObesity
NA
## condition_groupRenal failure
NA
## condition_groupSepsis
NA
## condition_groupVascular and unspecified dementia
NA
## age_group45-54
3.850
## regionNortheast
6.008
## regionSouth
5.358
## regionWest
5.603
##
t value
## (Intercept)
3.359
```

```
## conditionAdultRespDistSyn
-1.825
## conditionAll other conditions and causes (residual)
1.844
## conditionAlzheimer
-5.137
## conditionCardiacArrest
-2.118
## conditionCardiacArrhythmia
-3.990
## conditionCerebrovascular
-3.682
## conditionChronicLR
-3.754
## conditionCOVID-19
7.370
## conditionDiabetes
-1.753
## conditionHeartFailure
-3.655
## conditionHypertensive
-1.996
## conditionInfluenza
1.783
## conditioninjuryPoisoning
-3.943
## conditionIschemicheart
-3.462
## conditionMalignant
-3.763
## conditionObesity
-2.330
## conditionOtherCirculatorySystem
-3.359
## conditionOtherRespSystem
-3.663
## conditionRenal failure
-3.228
## conditionRespiratoryArrest
-3.934
## conditionSepsis
-2.656
## conditionVascular
-4.833
## condition_groupAll other conditions and causes (residual)
NA
## condition_groupAlzheimer disease
NA
## condition_groupCirculatory diseases
NA
```

```
## condition_groupCoronavirus Disease 2019
NA
## condition_groupDiabetes
NA
## condition_groupIntentional and unintentional injury, poisoning and other a
dverse events      NA
## condition_groupMalignant neoplasms
NA
## condition_groupObesity
NA
## condition_groupRenal failure
NA
## condition_groupSepsis
NA
## condition_groupVascular and unspecified dementia
NA
## age_group45-54
7.136
## regionNortheast
2.471
## regionSouth
3.016
## regionWest
2.188
##
Pr(>|t|)
## (Intercept)
0.000804
## conditionAdultRespDistSyn
0.068178
## conditionAll other conditions and causes (residual)
0.065355
## conditionAlzheimer
3.21e-07
## conditionCardiacArrest
0.034375
## conditionCardiacArrhythmia
6.95e-05
## conditionCerebrovascular
0.000240
## conditionChronicLR
0.000182
## conditionCOVID-19
2.98e-13
## conditionDiabetes
0.079896
## conditionHeartFailure
0.000267
## conditionHypertensive
0.046129
```

```
## conditionInfluenza
0.074749
## conditioninjuryPoisoning
8.47e-05
## conditionIschemicheart
0.000552
## conditionMalignant
0.000175
## conditionObesity
0.019960
## conditionOtherCirculatorySystem
0.000804
## conditionOtherRespSystem
0.000259
## conditionRenal failure
0.001275
## conditionRespiratoryArrest
8.79e-05
## conditionSepsis
0.008003
## conditionVascular
1.50e-06
## condition_groupAll other conditions and causes (residual)
NA
## condition_groupAlzheimer disease
NA
## condition_groupCirculatory diseases
NA
## condition_groupCoronavirus Disease 2019
NA
## condition_groupDiabetes
NA
## condition_groupIntentional and unintentional injury, poisoning and other a
dverse events      NA
## condition_groupMalignant neoplasms
NA
## condition_groupObesity
NA
## condition_groupRenal failure
NA
## condition_groupSepsis
NA
## condition_groupVascular and unspecified dementia
NA
## age_group45-54
1.57e-12
## regionNortheast
0.013580
## regionSouth
0.002607
```

```
## regionWest
0.028876
##
## (Intercept)
***
## conditionAdultRespDistSyn
.
## conditionAll other conditions and causes (residual)
.
## conditionAlzheimer
***  
## conditionCardiacArrest
*
## conditionCardiacArrhythmia
***  
## conditionCerebrovascular
***  
## conditionChronicLR
***  
## conditionCOVID-19
***  
## conditionDiabetes
.
## conditionHeartFailure
***  
## conditionHypertensive
*
## conditionInfluenza
.
## conditioninjuryPoisoning
***  
## conditionIschemicheart
***  
## conditionMalignant
***  
## conditionObesity
*
## conditionOtherCirculatorySystem
***  
## conditionOtherRespSystem
***  
## conditionRenal failure
**
## conditionRespiratoryArrest
***  
## conditionSepsis
**
## conditionVascular
***  
## condition_groupAll other conditions and causes (residual)
```

```
## condition_groupAlzheimer disease
## condition_groupCirculatory diseases
## condition_groupCoronavirus Disease 2019
## condition_groupDiabetes
## condition_groupIntentional and unintentional injury, poisoning and other a
dverse events
## condition_groupMalignant neoplasms
## condition_groupObesity
## condition_groupRenal failure
## condition_groupSepsis
## condition_groupVascular and unspecified dementia
## age_group45-54
***  

## regionNortheast
*  

## regionSouth
**  

## regionWest
*  

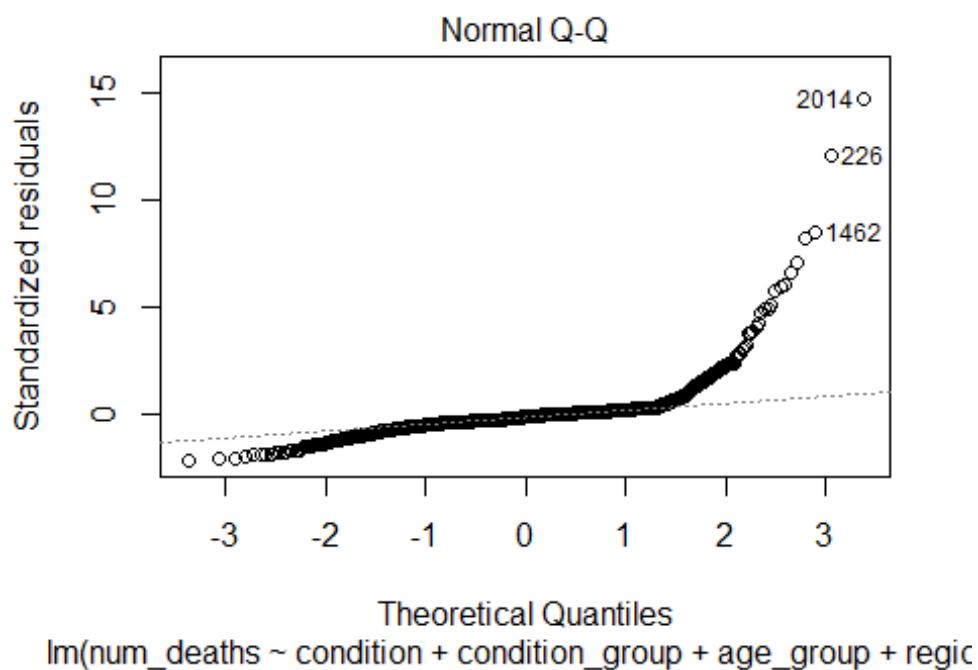
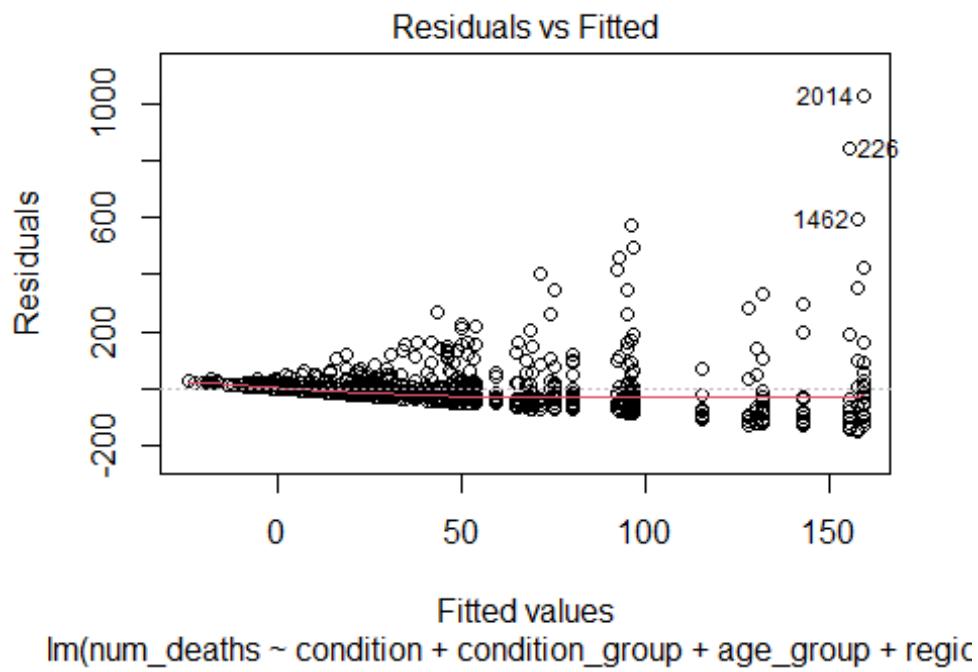
## ---  

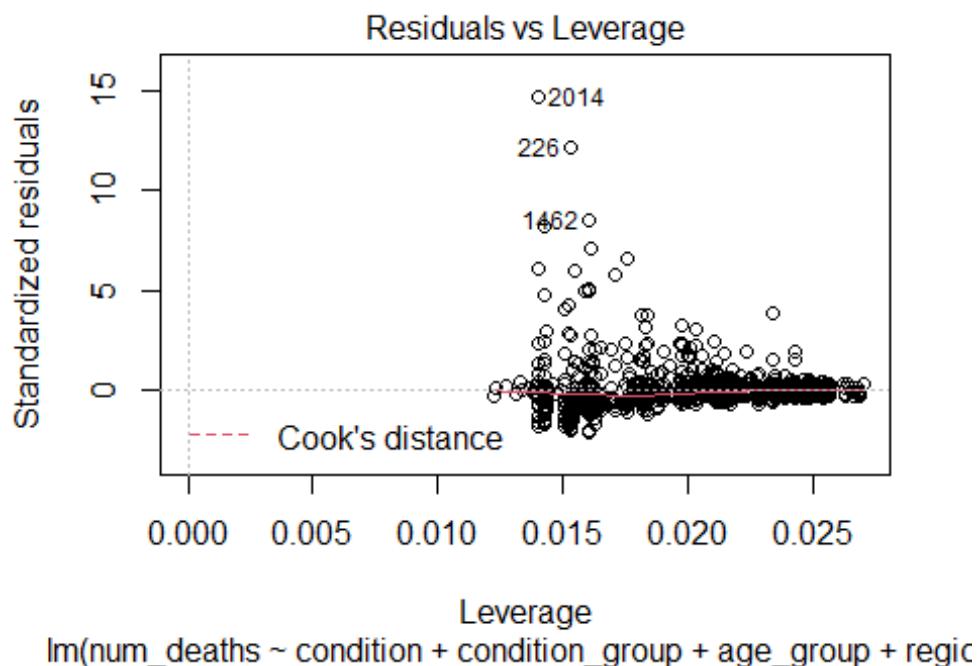
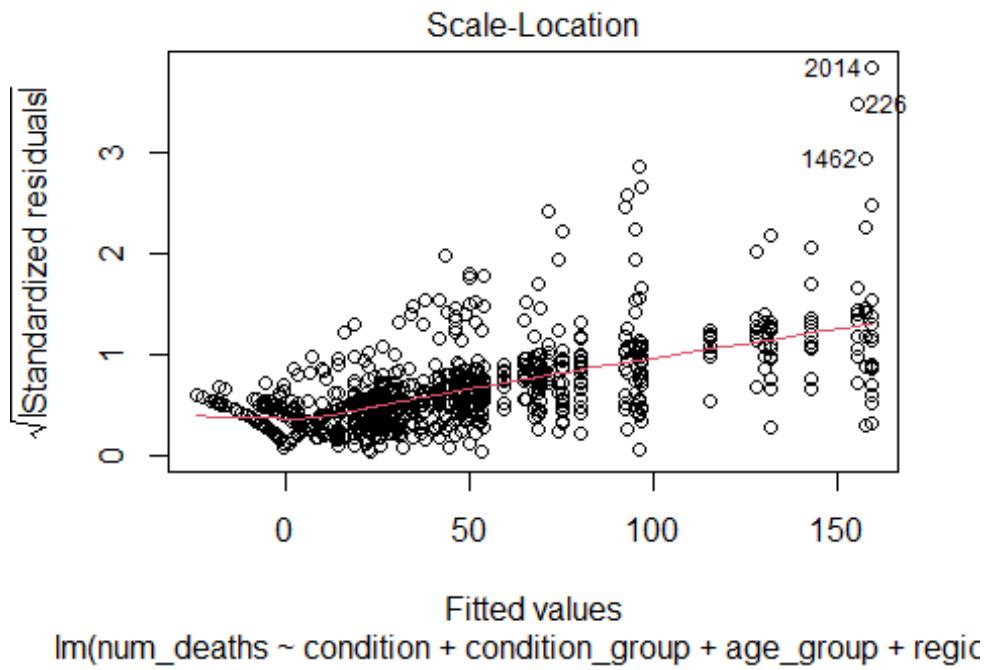
## Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  

##  

## Residual standard error: 70.33 on 1337 degrees of freedom
## Multiple R-squared: 0.249, Adjusted R-squared: 0.2344
## F-statistic: 17.05 on 26 and 1337 DF, p-value: < 2.2e-16  

lm_df <- tidy(lm_fit)
plot(lm_fit)
```





divide your data set into train and test data sets

```
set.seed(123)
partition <- sample(2, nrow(covidcleansed), replace = TRUE, prob= c(0.80,0.20))
train <- covidcleansed[partition==1, ]
```

```

test <- covidcleansed[partition==2, ]

library(MASS)
##data = covidcleanse
model.covid19_deathsA <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allo
thcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West, data=train)
summary(model.covid19_deathsA)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##   influenza + injuryPoisoning + ischemicheart + malignant +
##   obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##   conditionrespiratoryarrest + sepsis + vascular + Northeast +
##   South + West, data = train)
##
## Residuals:
##    Min      1Q      Median      3Q      Max
## -140.37  -22.71   -5.43     7.97 1033.47
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 37.334    10.725   3.481 0.000520 ***
## `age45-54`                  27.141     4.339   6.255 5.71e-10 ***
## adultrespdistSyn            -28.212    14.255  -1.979 0.048057 *
## allothcondsandcauses        4.854    13.063   0.372 0.710268
## alzheimer                  -61.371   12.443  -4.932 9.41e-07 ***
## cardiacArrest                -38.969   14.445  -2.698 0.007090 **
## cardiacarrhythmia           -57.393   13.867  -4.139 3.77e-05 ***
## cerebrovascula              -56.799   14.675  -3.871 0.000115 ***
## chroniclr                   -55.493   14.087  -3.939 8.70e-05 ***
## covid19                      70.129   12.773   5.490 5.01e-08 ***
## diabetes                     -27.196   13.793  -1.972 0.048890 *
## heartfailure                 -54.422   16.225  -3.354 0.000824 ***
## hypertensive                  -29.729   14.009  -2.122 0.034051 *
## influenza                     24.631   13.381   1.841 0.065942 .
## injuryPoisoning              -54.753   14.193  -3.858 0.000121 ***
## ischemicheart                 -50.605   14.476  -3.496 0.000492 ***
## malignant                     -54.328   15.172  -3.581 0.000358 ***
## obesity                       -32.189   14.360  -2.242 0.025189 *
## othercirculatorysystem       -50.783   14.891  -3.410 0.000673 ***

```

```

## otherrespstystem      -55.488    15.431   -3.596  0.000338 ***
## renalfailure          -50.578    14.366   -3.521  0.000448 ***
## conditionrespiratoryarrest -57.530    14.922   -3.855  0.000122 ***
## sepsis                 -37.822    13.725   -2.756  0.005955 **
## vascular               -58.878    12.895   -4.566  5.54e-06 ***
## Northeast              15.766     6.812    2.315  0.020827 *
## South                  16.923     6.014    2.814  0.004987 **
## West                   12.230     6.315    1.937  0.053028 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.72 on 1075 degrees of freedom
## Multiple R-squared:  0.2394, Adjusted R-squared:  0.221
## F-statistic: 13.01 on 26 and 1075 DF,  p-value: < 2.2e-16

prediction <- predict(model.covid19_deathsA,test)
actual =test$num_deaths
cor(prediction, actual)

## [1] 0.5258685

model.covid19_deathsB <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allo
thcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespstystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West, data=train)
summary(model.covid19_deathsB)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + heartfailure + hypertensive + influenza +
##   injuryPoisoning + ischemicheart + malignant + obesity + othercirculato
rysystem +
##   otherrespstystem + renalfailure + conditionrespiratoryarrest +
##   sepsis + vascular + Northeast + South + West, data = train)
##
## Residuals:
##       Min     1Q   Median     3Q    Max
## -139.80  -22.57   -5.44    8.59 1033.61
##
## Coefficients:
## (Intercept)           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           24.645     8.592   2.868  0.004206 **
## `age45-54`            26.927     4.343   6.200  8.04e-10 ***
## adultrespdistSyn     -15.400    12.705  -1.212  0.225726

```

```

## allothcondsandcauses      17.667    11.348   1.557  0.119789
## alzheimer                 -48.604   10.640   -4.568  5.49e-06 ***
## cardiacArrest              -26.170   12.922   -2.025  0.043086 *
## cardiacarrhythmia          -44.650   12.286   -3.634  0.000292 ***
## cerebrovascula             -44.045   13.190   -3.339  0.000869 ***
## chroniclr                  -42.701   12.521   -3.410  0.000673 ***
## covid19                     82.905   11.023   7.521  1.14e-13 ***
## heartfailure                -41.747   14.918   -2.798  0.005226 **
## hypertensive                -16.886   12.419   -1.360  0.174217
## influenza                   37.419   11.720   3.193  0.001450 **
## injuryPoisoning            -41.941   12.636   -3.319  0.000933 ***
## ischemicheart               -37.845   12.967   -2.919  0.003590 **
## malignant                   -41.537   13.734   -3.024  0.002550 **
## obesity                      -19.404   12.830   -1.512  0.130720
## othercirculatorysystem     -38.006   13.426   -2.831  0.004728 **
## otherrespsystem              -42.760   14.035   -3.047  0.002370 **
## renalfailure                -37.763   12.829   -2.944  0.003314 **
## conditionrespiratoryarrest -44.714   13.450   -3.324  0.000916 ***
## sepsis                       -24.961   12.092   -2.064  0.039236 *
## vascular                     -46.134   11.173   -4.129  3.92e-05 ***
## Northeast                    15.325    6.817    2.248  0.024781 *
## South                        16.915    6.022    2.809  0.005065 **
## West                         12.590    6.321    1.992  0.046639 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.81 on 1076 degrees of freedom
## Multiple R-squared:  0.2366, Adjusted R-squared:  0.2189
## F-statistic: 13.34 on 25 and 1076 DF,  p-value: < 2.2e-16

prediction <- predict(model.covid19_deathsB,test)
actual =test$num_deaths
cor(prediction, actual)

## [1] 0.5299681

#plot(prediction,actual)

```

5 fold cross validation to validate our model

```

library(DAAG)

## Loading required package: lattice

##
## Attaching package: 'DAAG'

## The following object is masked from 'package:MASS':
##
##      hills

```

```

outA <- cv.lm(data = test
              , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespssystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West)
              , plotit = "Observed", m=5)

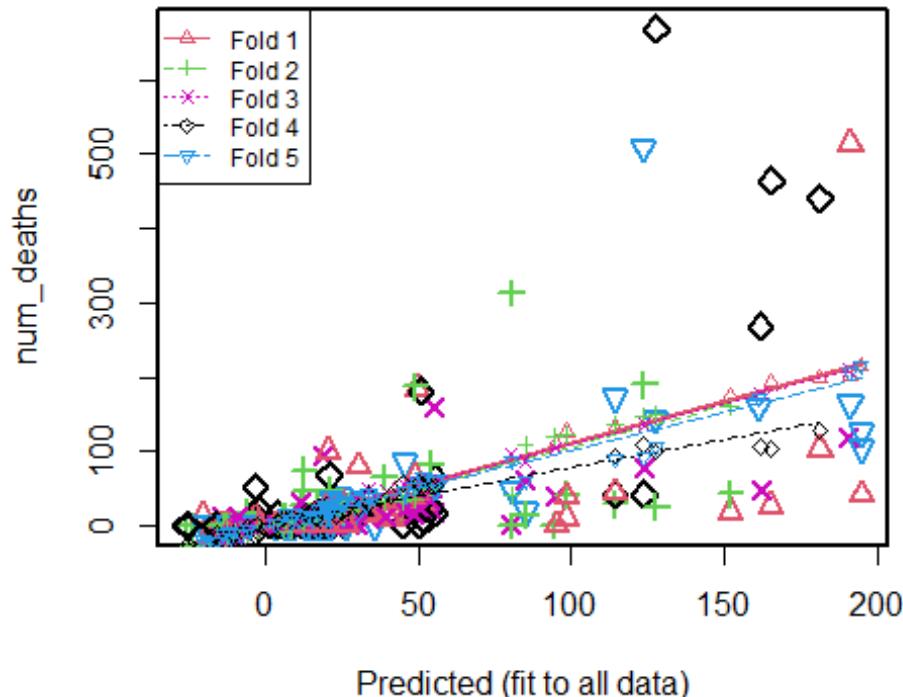
## Analysis of Variance Table
##
## Response: num_deaths
##                                     Df  Sum Sq Mean Sq F value    Pr(>F)
## `age45-54`                         1  55638  55638  11.56 0.00079 ***
## adultrespdistSyn                   1     308     308   0.06 0.80048
## allothcondsandcauses              1 124916 124916 25.95 7.2e-07 ***
## alzheimer                          1  17801  17801   3.70 0.05568 .
## cardiacArrest                      1    9110    9110   1.89 0.17021
## cardiacarrhythmia                 1    1509    1509   0.31 0.57603
## cerebrovascula                     1     512     512   0.11 0.74459
## chroniclr                          1    1137    1137   0.24 0.62739
## covid19                            1 355946 355946 73.95 1.1e-15 ***
## diabetes                           1    3533    3533   0.73 0.39249
## heartfailure                       1     499     499   0.10 0.74768
## hypertensive                        1     196     196   0.04 0.84014
## influenza                           1    7406    7406   1.54 0.21606
## injuryPoisoning                    1     992     992   0.21 0.65025
## ischemicheart                      1     500     500   0.10 0.74761
## malignant                           1    1759    1759   0.37 0.54612
## obesity                             1      5      5   0.00 0.97309
## othercirculatorysystem             1     46      46   0.01 0.92223
## otherrespssystem                    1    203    203   0.04 0.83740
## renalfailure                        1   4668   4668   0.97 0.32575
## conditionrespiratoryarrest        1   1303   1303   0.27 0.60332
## sepsis                             1    176    176   0.04 0.84870
## vascular                           1 12311 12311   2.56 0.11111
## Northeast                          1    113    113   0.02 0.87819
## South                             1   3229   3229   0.67 0.41356
## West                              1   2523   2523   0.52 0.46983
## Residuals                          235 1131151 4813
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
: 
##
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function

```

```
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate
```

### Small symbols show cross-validation predicted val



```
##
## fold 1
## Observations in test set: 52
##          21    27    38     79   186    280   359    396    413    418
499
## Predicted  3.37  18.9  18.0 -12.60  49.8  50.83  165   17.5  25.00  27.17  2
2.0
## cvpred    6.24  18.2  11.9 -2.63  36.3  41.12  192   21.2  32.52  30.48  3
0.4
## num_deaths 0.00  0.0   0.0   0.00  185.0  39.00   26   0.0  30.00  21.00  0.0
0.0
## CV residual -6.24 -18.2 -11.9   2.63  148.7 -2.12 -166 -21.2 -2.52 -9.48 -3
0.4
##          509   544   578   609   668   675   699   726   728   810    8
18
## Predicted  6.13  41.9  4.05 -7.86  20.3 -20.7 -13.2  181.2  41.9  16.7  1
95
## cvpred    5.67  48.0  3.11 -7.02  12.1 -17.1 -14.9  197.7  48.0  20.7  2
16
## num_deaths 0.00  22.0  0.00  0.00  100.0  13.0   0.0  103.0  10.0   0.0
41
## CV residual -5.67 -26.0 -3.11  7.02  87.9  30.1  14.9 -94.7 -38.0 -20.7 -1
75
##          917   972  1007  1015  1100  1127  1139  1174  1266  1413
```

```

1464
## Predicted 25.00 19.7 11.1 -3.2350 40.4 -9.04 152 33.7 33.7 94.8
51.3
## cvpred 32.52 14.6 7.3 -0.0395 30.6 -11.17 174 30.2 30.2 112.0
53.7
## num_deaths 24.00 0.0 14.0 11.0000 11.0 0.00 18 17.0 14.0 0.0
31.0
## CV residual -8.52 -14.6 6.7 11.0395 -19.6 11.17 -156 -13.2 -16.2 -112.0
-22.7
## 1500 1517 1564 1586 1589 1600 1689 1892 1909 1914
1965
## Predicted 12.69 -11.16 18.71 48.7 -15.4 191 98.3 9.82 -9.04 3.32
98.3
## cvpred 8.29 -9.67 11.31 39.2 -16.8 201 126.7 11.25 -11.17 2.57
126.7
## num_deaths 0.00 0.00 14.00 64.0 0.0 514 10.0 0.00 0.00 0.00
39.0
## CV residual -8.29 9.67 2.69 24.8 16.8 313 -116.7 -11.25 11.17 -2.57
-87.7
## 1994 2071 2138 2174 2236 2275 2304 2313
## Predicted 30.5 -2.17 14.0 20.93 114.2 19.6 26.5 15.5
## cvpred 29.6 7.17 6.5 17.31 131.9 31.3 17.3 18.9
## num_deaths 79.0 0.00 0.0 15.00 44.0 0.0 0.0 0.0
## CV residual 49.4 -7.17 -6.5 -2.31 -87.9 -31.3 -17.3 -18.9
##
## Sum of squares = 274176      Mean square = 5273      n = 52
##
## fold 2
## Observations in test set: 53
## 39 125 156 178 198 266 310 352 355 371
397
## Predicted 94.3 -12.6 21.48 123.6 80.3 44.85 13.2 16.7 -7.3 -2.168 -
12.6
## cvpred 120.3 -15.0 17.82 148.2 40.7 31.59 12.2 12.9 -6.0 -0.861 -
15.0
## num_deaths 0.0 0.0 11.00 190.0 313.0 22.00 0.0 0.0 0.0 0.000
0.0
## CV residual -120.3 15.0 -6.82 41.8 272.3 -9.59 -12.2 -12.9 6.0 0.861
15.0
## 412 457 458 492 503 517 571 586 635 656 7
36
## Predicted 53.8 24.47 53.81 48.90 20.4 -7.86 -23.7 114 12.60 12.11 8.
81
## cvpred 50.1 22.19 50.08 34.33 19.5 -10.06 -28.3 137 4.36 6.98 8.
80
## num_deaths 82.0 15.00 53.00 42.00 0.0 0.00 0.0 29 74.00 28.00 0.
00
## CV residual 31.9 -7.19 2.92 7.67 -19.5 10.06 28.3 -108 69.64 21.02 -8.
80
## 876 894 900 977 982 1003 1028 1034 1039 1041 10

```

```

84
## Predicted    23.7 15.57 14.0 -7.33 27.0 12.60 17.10 38.8 -26.0 6.14 4.
05
## cvpred      26.4 12.46 16.1 -8.00 32.1 4.36 19.23 33.4 -28.5 -7.13 3.
21
## num_deaths   10.0 11.00 0.0 0.00 0.0 46.00 25.00 65.0 0.0 12.00 0.
00
## CV residual -16.4 -1.46 -16.1 8.00 -32.1 41.64 5.77 31.6 28.5 19.13 -3.
21
##          1204 1222 1239 1288 1321 1511 1559 1567 1645 1675 1690
## Predicted    80.3 13.4 25.0 8.81 84.9 3.87 21.0 6.67 152 -10.2 128
## cvpred       40.7 14.0 25.2 8.80 109.5 5.51 21.5 8.25 159 -14.8 151
## num_deaths   0.0 0.0 14.0 0.00 14.0 0.00 47.0 12.00 43 0.0 25
## CV residual -40.7 -14.0 -11.2 -8.80 -95.5 -5.51 25.5 3.75 -116 14.8 -126
##          1734 1773 1873 1983 2008 2035 2061 2097 2211
## Predicted    18.0 -15.4 98.3 -6.1 48.9 -7.86 26.02 -12.6 -13.2
## cvpred       19.2 -11.8 123.0 -10.6 34.3 -10.06 17.93 -15.0 -18.1
## num_deaths   0.0 0.0 42.0 17.0 187.0 0.00 20.00 0.0 0.0
## CV residual -19.2 11.8 -81.0 27.6 152.7 10.06 2.07 15.0 18.1
##
## Sum of squares = 189698      Mean square = 3579      n = 53
##
## fold 3
## Observations in test set: 53
##          16 190 209 211 274 307 325 428 454 487 4
91
## Predicted    80.3 18.2 -14.3 18.9 51.8 -15.4 25.0 25.5 55.4 -11.9 19.
56
## cvpred       97.4 25.7 -19.8 15.8 56.0 -15.1 22.4 27.3 53.8 -17.2 19.
89
## num_deaths   0.0 14.0 13.0 93.0 23.0 0.0 0.0 17.0 158.0 0.0 12.
00
## CV residual -97.4 -11.7 32.8 77.2 -33.0 15.1 -22.4 -10.3 104.2 17.2 -7.
89
##          527 597 677 774 793 868 904 907 909 967
994
## Predicted    18.9 -11.16 84.9 55.4 -3.81 33.7 45.4 94.8 162 -5.69 1
3.2
## cvpred       15.8 -8.76 87.0 53.8 -7.15 48.4 56.5 104.3 178 -1.41 2
2.4
## num_deaths   0.0 0.00 61.0 18.0 0.00 10.0 21.0 40.0 48 0.00
0.0
## CV residual -15.8 8.76 -26.0 -35.8 7.15 -38.4 -35.5 -64.3 -130 1.41 -2
2.4
##          1009 1014 1071 1213 1225 1327 1334 1375 1391 1438
1472
## Predicted    11.58 13.753 -6.0 -8.42 15.5 -5.5 8.81 21 -7.33 22.0
18.2
## cvpred       7.24 15.736 -14.5 -18.11 19.0 -3.4 11.41 27 -5.02 29.5
25.7

```

```

## num_deaths 32.00 15.000 0.0 0.00 0.0 0.0 0.00 0 0.00 13.0
0.0
## CV residual 24.76 -0.736 14.5 18.11 -19.0 3.4 -11.41 -27 5.02 -16.5 -
25.7
## 1523 1540 1554 1760 1805 1818 1836 1838 1936 1960
2041
## Predicted -10.16 48.2 190.6 22.0 -7.33 30.2 53.8 54.35 39.5 16.7
-8.42
## cvpred -6.82 50.3 209.3 29.5 -5.02 37.3 59.4 56.91 45.9 20.3
-18.11
## num_deaths 0.00 15.0 119.0 11.0 0.00 0.0 33.0 49.00 13.0 0.0
0.00
## CV residual 6.82 -35.3 -90.3 -18.5 5.02 -37.3 -26.4 -7.91 -32.9 -20.3
18.11
## 2050 2052 2125 2152 2190 2225 2248 2252 2255
## Predicted 13.4 12.7 3.91 49.8 123.6 -9.04 27.2 23.2 0.254
## cvpred 16.3 19.4 2.79 58.5 135.8 -14.53 30.9 28.6 -2.070
## num_deaths 0.0 0.0 0.00 25.0 78.0 13.00 0.0 0.0 0.000
## CV residual -16.3 -19.4 -2.79 -33.5 -57.8 27.53 -30.9 -28.6 2.070
##
## Sum of squares = 81147 Mean square = 1531 n = 53
##
## fold 4
## Observations in test set: 52
## 7 94 126 159 183 217 313 320 382 385 4
34
## Predicted 21.0 53.8 16.7 -2.88 3.87 -16.7 -10.8 55.4 25.5 1.18 24.
98
## cvpred 24.3 53.3 10.4 1.56 5.39 -12.2 -10.9 53.1 20.7 -7.26 17.
15
## num_deaths 0.0 23.0 0.0 10.00 15.00 0.0 0.0 63.0 0.0 0.00 25.
00
## CV residual -24.3 -30.3 -10.4 8.44 9.61 12.2 10.9 9.9 -20.7 7.26 7.
85
## 550 628 630 647 680 761 808 881 936 944 96
2
## Predicted 40.9 44.8 18.0 -3.24 181 -25.3 17.5 51.5 30.5 33.72 50.
3
## cvpred 41.2 57.4 25.3 -13.61 128 -25.7 11.5 53.9 24.2 28.79 57.
5
## num_deaths 21.0 0.0 0.0 52.00 440 0.0 0.0 10.0 17.0 33.00 0.
0
## CV residual -20.2 -57.4 -25.3 65.61 312 25.7 -11.5 -43.9 -7.2 4.21 -57.
5
## 963 965 992 1083 1136 1138 1171 1197 1221 1308
1315
## Predicted 21.5 -10.6 14.0 -25.3 8.62 114.2 22.91 -6.22 -15.9 5.67
-26.0
## cvpred 19.6 -15.5 15.7 -25.7 10.69 93.7 14.29 -2.30 -11.1 11.50
-26.8

```

```

## num_deaths      0.0    0.0    0.0    0.0    0.00   42.0  12.00   0.00    0.0    0.00
0.0
## CV residual -19.6   15.5 -15.7   25.7 -10.69 -51.7 -2.29   2.30   11.1 -11.50
26.8
##          1421  1461  1490  1506  1513  1544  1562  1626  1636  1648  177
5
## Predicted     21    162   20.9 123.6   20.4   13.4   50.8  17.10   4.05   55.4 -16.
1
## cvpred       26    107   26.0 108.2   30.7   20.4   51.1  18.44   5.77   53.1 -16.
9
## num_deaths    67    268   10.0  42.0   15.0    0.0 180.0  16.00   0.00   17.0    0.
0
## CV residual   41    161  -16.0 -66.2 -15.7 -20.4 128.9 -2.44 -5.77 -36.1  16.
9
##          1776  1814  1860  1883   1979  2012  2013  2019
## Predicted    13.2   15.6   19.1 11.58  -2.17 127.7  165   20.4
## cvpred       14.6   21.4   17.3  9.72 -11.12  99.4  103   30.7
## num_deaths    0.0    0.0    0.0   0.00   22.00 668.0  464    0.0
## CV residual  -14.6  -21.4  -17.3 -9.72  33.12 568.6  361  -30.7
##
## Sum of squares = 627860      Mean square = 12074      n = 52
##
## fold 5
## Observations in test set: 52
##          5    12    26    31   123   152   176   179   204   224
306
## Predicted   20.4   35.5   15.0 -15.9 -11.87  80.3  17.98 161.2 32.7 123.6  3
0.2
## cvpred      15.9   38.4   11.3 -20.9  -7.61  86.4  12.57 172.4 28.1  93.5  3
5.6
## num_deaths   0.0    0.0    0.0    0.0    0.00   48.0  15.00 160.0 34.0 509.0  1
0.0
## CV residual -15.9  -38.4  -11.3   20.9   7.61 -38.4   2.43 -12.4   5.9 415.5 -2
5.6
##          411   486   533   537   757   840   849   858   871   910
946
## Predicted   24.5   33.7  -16.7 -11.4   9.49   25.5   22.9  48.9 21.49 191.1  1
7.5
## cvpred      29.2   42.5  -22.8 -16.4  16.92   29.8   34.3  51.8 23.06 207.7  2
1.3
## num_deaths  40.0   25.0    0.0    0.0    0.00   17.0   18.0  28.0 22.00 164.0
0.0
## CV residual 10.8  -17.5   22.8  16.4 -16.92 -12.8  -16.3 -23.8 -1.06 -43.7 -2
1.3
##          949   954   985  1046  1090  1091  1207  1216  1287  1293
1312
## Predicted   19.56 127.7 -13.8 114.2   8.62   84.9  -7.86  15.0 -20.5 -19.5
20.3
## cvpred     22.89 106.8 -11.2  89.5   8.57   60.6 -12.45 11.3 -20.3 -23.8
25.2

```



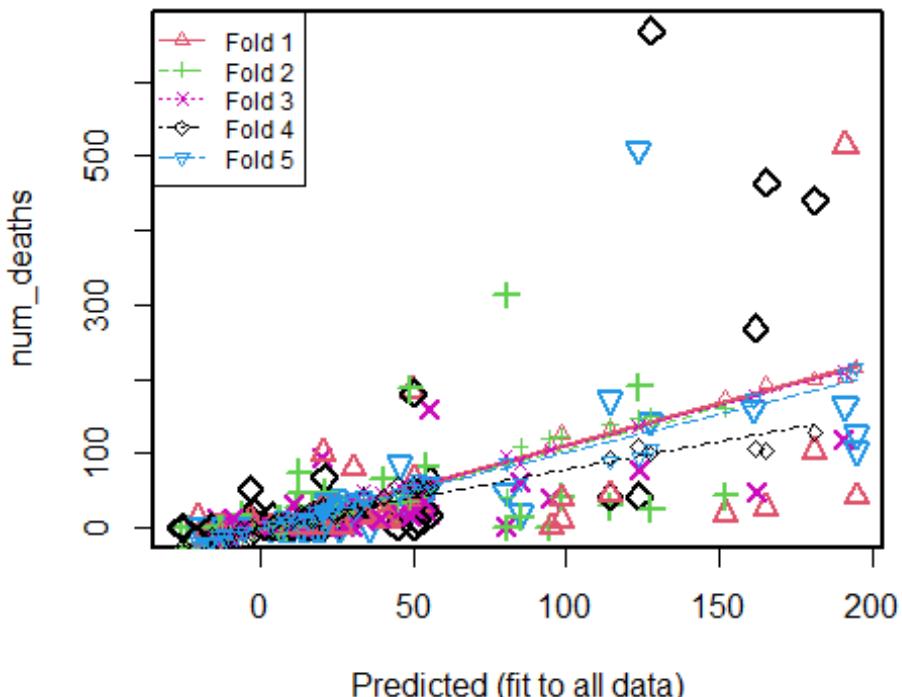
```

## heartfailure           1    742    742   0.15 0.69435
## hypertensive          1    102    102   0.02 0.88441
## influenza              1   6584   6584   1.37 0.24238
## injuryPoisoning       1   1354   1354   0.28 0.59557
## ischemicheart         1    711    711   0.15 0.70042
## malignant              1   2248   2248   0.47 0.49409
## obesity                1     71     71   0.01 0.90307
## othercirculatorysystem 1    152    152   0.03 0.85884
## otherrespsystem        1    461    461   0.10 0.75674
## renalfailure           1   3340   3340   0.70 0.40472
## conditionrespiratoryarrest 1   2272   2272   0.47 0.49180
## sepsis                 1    483    483   0.10 0.75131
## vascular               1 15025  15025   3.13 0.07794 .
## Northeast              1    117    117   0.02 0.87587
## South                  1   3273   3273   0.68 0.40941
## West                   1   2501   2501   0.52 0.47078
## Residuals             236 1131177 4793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
:
## 
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

## Small symbols show cross-validation predicted val



```

## 
## fold 1
## Observations in test set: 52
##          21     27     38     79    186    280    359    396    413    418
499
## Predicted   3.34   20.1   17.9  -12.57   49.7  50.00   165   17.5  24.20  27.19   2
1.9
## cvpred     6.21   19.3   11.8  -2.62   36.3  40.36   192   21.2  31.75  30.48   3
0.4
## num_deaths  0.00    0.0    0.0    0.00  185.0  39.00    26    0.0  30.00  21.00
0.0
## CV residual -6.21  -19.3  -11.8    2.62  148.7  -1.36  -166  -21.2  -1.75  -9.48  -3
0.4
##          509    544    578    609    668    675    699    726    728    810    8
18
## Predicted   6.09   41.9   4.05  -7.91   20.3  -20.7  -13.2  181.2   41.9  16.8   1
95
## cvpred     5.63   48.0   3.11  -7.06   12.1  -17.1  -14.9  197.7   48.0  20.7   2
16
## num_deaths  0.00   22.0   0.00   0.00  100.0   13.0    0.0  103.0   10.0   0.0
41
## CV residual -5.63  -26.0  -3.11   7.06   87.9   30.1   14.9  -94.7  -38.0  -20.7  -1
75
##          917    972  1007   1015   1100   1127   1139   1174   1266   1413
1464
## Predicted   24.20   19.7  11.1  -3.2259   40.4  -9.05   152   33.7  33.7  94.8
51.2
## cvpred     31.75   14.7   7.3  -0.0453   30.6  -11.18   174   30.2  30.2  112.0
53.7
## num_deaths  24.00    0.0  14.0  11.0000   11.0    0.00    18   17.0  14.0   0.0
31.0
## CV residual -7.75  -14.7   6.7  11.0453  -19.6   11.18  -156  -13.2  -16.2  -112.0
-22.7
##          1500   1517  1564  1586   1589   1600   1689   1892   1909   1914
1965
## Predicted   12.62  -11.20  18.74  50.0  -15.4   191   98.4   9.81  -9.05  3.31
98.4
## cvpred     8.24   -9.69  11.35  40.4  -16.8   201   126.7  11.25  -11.18  2.57
126.7
## num_deaths  0.00    0.00  14.00  64.0    0.0   514   10.0    0.00   0.00   0.00
39.0
## CV residual -8.24    9.69   2.65  23.6   16.8   313  -116.7  -11.25  11.18  -2.57
-87.7
##          1994   2071  2138  2174  2236  2275  2304  2313
## Predicted   30.6  -2.14  13.96  20.90  114.2  19.6  26.4  15.5
## cvpred     29.7   7.18   6.52  17.29  131.9  31.3  17.2  18.9
## num_deaths  79.0   0.00   0.00  15.00  44.0    0.0   0.0   0.0
## CV residual 49.3  -7.18  -6.52  -2.29  -87.9  -31.3  -17.2  -18.9
##
## Sum of squares = 274156      Mean square = 5272      n = 52

```

```

## 
## fold 2
## Observations in test set: 53
##          39   125   156   178   198   266   310   352   355   371
397
## Predicted    94.2 -12.6 21.42 124  80.2 44.79 13.2 16.8 -7.24 -2.144 -1
2.6
## cvpred     120.1 -14.9 17.63 148  40.6 31.43 12.3 13.0 -5.89 -0.832 -1
4.9
## num_deaths   0.0   0.0 11.00 190 313.0 22.00   0.0   0.0  0.00  0.000
0.0
## CV residual -120.1 14.9 -6.63  42 272.4 -9.43 -12.3 -13.0  5.89  0.832  1
4.9
##          412   457   458   492   503   517   571   586   635   656   736
## Predicted   53.8 24.5 53.85 48.93 20.4 -7.91 -23.7 114 12.6 12.11 8.82
## cvpred      50.2 22.3 50.21 34.45 19.3 -10.29 -28.2 137 4.4 7.08 8.97
## num_deaths  82.0 15.0 53.00 42.00  0.0   0.00  0.0   29 74.0 28.00 0.00
## CV residual 31.8 -7.3  2.79  7.55 -19.3 10.29 28.2 -108 69.6 20.92 -8.97
##          876   894   900   977   982 1003 1028 1034 1039 1041 108
4
## Predicted   23.6 15.59 14.0 -7.30 27.0 12.6 17.09 40.1 -26.0  6.14  4.0
5
## cvpred      26.4 12.56 16.2 -7.98 32.2  4.4 19.34 37.3 -28.5 -7.03  3.3
0
## num_deaths  10.0 11.00  0.0  0.00  0.0 46.0 25.00 65.0   0.0 12.00  0.0
0
## CV residual -16.4 -1.56 -16.2  7.98 -32.2 41.6  5.66 27.7  28.5 19.03 -3.3
0
##          1204  1222  1239  1288  1321  1511 1559 1567 1645 1675 1690
## Predicted   80.2 13.4 24.20 8.82 84.9 3.83 21.0 6.69 152 -10.2 128
## cvpred      40.6 13.8 22.92 8.97 109.5 5.29 21.6 8.30 159 -14.6 151
## num_deaths   0.0   0.0 14.00  0.00 14.0  0.00 47.0 12.00  43   0.0  25
## CV residual -40.6 -13.8 -8.92 -8.97 -95.5 -5.29 25.4  3.70 -116 14.6 -126
##          1734  1773  1873  1983  2008  2035 2061 2097 2211
## Predicted   17.9 -15.4 98.4 -6.07 48.9 -7.91 26.05 -12.6 -13.2
## cvpred      19.0 -11.8 123.1 -10.58 34.4 -10.29 17.97 -14.9 -18.1
## num_deaths   0.0   0.0 42.0  17.00 187.0  0.00 20.00   0.0   0.0
## CV residual -19.0 11.8 -81.1 27.58 152.6 10.29  2.03 14.9  18.1
##
## Sum of squares = 189407    Mean square = 3574    n = 53
##
## fold 3
## Observations in test set: 53
##          16   190   209   211   274   307   325   428   454   487   4
91
## Predicted   80.2 18.1 -14.3 20.1 51.8 -15.4 24.2 25.6 55.4 -11.8 19.
60
## cvpred     97.4 25.7 -19.8 19.5 56.0 -15.1 20.4 27.3 53.9 -17.2 19.
91
## num_deaths   0.0 14.0 13.0 93.0 23.0   0.0   0.0 17.0 158.0   0.0 12.

```

```

00
## CV residual -97.4 -11.7 32.8 73.5 -33.0 15.1 -20.4 -10.3 104.1 17.2 -7.
91
## 527 597 677 774 793 868 904 907 909 967
994
## Predicted 20.1 -11.20 84.9 55.4 -3.76 33.8 45.4 94.8 162 -5.68 1
3.2
## cvpred 19.5 -8.72 86.9 53.9 -7.13 48.4 56.5 104.4 178 -1.39 2
2.4
## num_deaths 0.0 0.00 61.0 18.0 0.00 10.0 21.0 40.0 48 0.00
0.0
## CV residual -19.5 8.72 -25.9 -35.9 7.13 -38.4 -35.5 -64.4 -130 1.39 -2
2.4
## 1009 1014 1071 1213 1225 1327 1334 1375 1391 143
8
## Predicted 10.75 13.731 -5.97 -8.43 15.5 -5.48 8.82 21.0 -7.30 22.
0
## cvpred 5.07 15.638 -14.61 -18.09 19.0 -3.44 11.36 27.1 -4.99 29.
5
## num_deaths 32.00 15.000 0.00 0.00 0.0 0.00 0.00 0.0 0.00 13.
0
## CV residual 26.93 -0.638 14.61 18.09 -19.0 3.44 -11.36 -27.1 4.99 -16.
5
## 1472 1523 1540 1554 1760 1805 1818 1836 1838 1936
1960
## Predicted 18.1 -10.21 49.4 190.5 22.0 -7.30 30.2 53.8 53.53 39.6
16.8
## cvpred 25.7 -6.77 53.9 209.4 29.5 -4.99 37.3 59.4 54.81 45.9
20.3
## num_deaths 0.0 0.00 15.0 119.0 11.0 0.00 0.0 33.0 49.00 13.0
0.0
## CV residual -25.7 6.77 -38.9 -90.4 -18.5 4.99 -37.3 -26.4 -5.81 -32.9 -
20.3
## 2041 2050 2052 2125 2152 2190 2225 2248 2252 2255
## Predicted -8.43 13.4 12.6 3.95 49.7 123.6 -9.05 27.2 23.3 0.28
## cvpred -18.09 16.3 19.4 2.82 58.5 135.8 -14.56 30.9 28.6 -2.01
## num_deaths 0.00 0.0 0.0 0.00 25.0 78.0 13.00 0.0 0.0 0.00
## CV residual 18.09 -16.3 -19.4 -2.82 -33.5 -57.8 27.56 -30.9 -28.6 2.01
##
## Sum of squares = 81014 Mean square = 1529 n = 53
##
## fold 4
## Observations in test set: 52
## 7 94 126 159 183 217 313 320 382 385
434
## Predicted 20.1 53.8 16.8 -2.93 3.83 -16.7 -10.8 55.38 25.6 1.22 25
.05
## cvpred 23.7 53.4 10.5 1.54 5.37 -12.2 -10.8 53.14 20.7 -7.19 17
.23
## num_deaths 0.0 23.0 0.0 10.00 15.00 0.0 0.0 63.00 0.0 0.00 25

```

```

.00
## CV residual -23.7 -30.4 -10.5  8.46  9.63  12.2  10.8  9.86 -20.7  7.19  7
.77
##          550   628   630   647   680   761   808   881   936   944   9
62
## Predicted 40.1  44.8  17.9 -3.23  181  -25.3  17.5  51.5 30.55 33.74  50
.3
## cvpred    40.6  57.3  25.2 -13.61 128  -25.7  11.6  53.9 24.25 28.83  57
.5
## num_deaths 21.0   0.0   0.0  52.00 440   0.0   0.0  10.0 17.00 33.00   0
.0
## CV residual -19.6 -57.3 -25.2  65.61 312  25.7 -11.6 -43.9 -7.25  4.17 -57
.5
##          963   965   992  1083  1136  1138  1171  1197  1221  1308
1315
## Predicted 20.7 -10.6  14.0 -25.3   8.63 114.2 24.20 -6.29 -16.0   5.67
-26.0
## cvpred    19.1 -15.5  15.7 -25.7  10.68  93.6 14.96 -2.35 -11.1  11.49
-26.8
## num_deaths  0.0   0.0   0.0   0.0   0.00  42.0 12.00  0.00   0.0   0.00
0.0
## CV residual -19.1  15.5 -15.7  25.7 -10.68 -51.6 -2.96  2.35  11.1 -11.49
26.8
##          1421  1461  1490  1506  1513  1544  1562  1626  1636  1648  177
5
## Predicted  21   162   20.9 123.6  20.4   13.4  50.0 17.09  4.05  55.4 -16.
1
## cvpred    26   107   26.0 108.2  30.6   20.3  50.5 18.43  5.74  53.1 -16.
9
## num_deaths 67   268   10.0  42.0  15.0   0.0 180.0 16.00  0.00  17.0   0.
0
## CV residual 41   161  -16.0 -66.2 -15.6 -20.3 129.5 -2.43 -5.74 -36.1  16.
9
##          1776  1814  1860  1883   1979  2012  2013  2019
## Predicted 13.2  15.6  19.1 10.75  -2.14 127.7  165  20.4
## cvpred    14.6  21.4  17.3  9.14 -11.08  99.5  103  30.6
## num_deaths  0.0   0.0   0.0   0.00  22.00 668.0  464   0.0
## CV residual -14.6 -21.4 -17.3 -9.14  33.08 568.5  361  -30.6
##
## Sum of squares = 627843      Mean square = 12074      n = 52
##
## fold 5
## Observations in test set: 52
##          5    12    26    31    123   152    176    179    204    224
306
## Predicted 20.4  35.4  15.0 -16.0 -11.83  80.2 17.94 161.2 32.68 123.6
30.2
## cvpred    16.0  38.6  11.6 -20.8  -7.73  86.4 12.74 172.6 28.32  93.7
35.7
## num_deaths  0.0   0.0   0.0   0.0   0.00  48.0 15.00 160.0 34.00 509.0

```

```

10.0
## CV residual -16.0 -38.6 -11.6 20.8 7.73 -38.4 2.26 -12.6 5.68 415.3 -
25.7
## 411 486 533 537 757 840 849 858 871 910 9
46
## Predicted 24.5 33.7 -16.7 -11.4 10.7 25.6 24.2 48.9 20.66 191.1 17
.5
## cvpred 29.1 42.5 -22.6 -16.2 14.1 29.7 31.4 51.7 24.59 207.8 21
.2
## num_deaths 40.0 25.0 0.0 0.0 0.0 17.0 18.0 28.0 22.00 164.0 0
.0
## CV residual 10.9 -17.5 22.6 16.2 -14.1 -12.7 -13.4 -23.7 -2.59 -43.8 -21
.2
## 949 954 985 1046 1090 1091 1207 1216 1287 1293
1312
## Predicted 19.60 127.7 -13.7 114.2 8.63 84.9 -7.91 15.0 -20.5 -19.5
20.3
## cvpred 22.74 106.7 -11.2 89.5 8.49 60.5 -12.29 11.6 -20.4 -23.8
25.2
## num_deaths 14.00 143.0 0.0 172.0 0.00 20.0 0.00 0.0 0.0 0.0
0.0
## CV residual -8.74 36.3 11.2 82.5 -8.49 -40.5 12.29 -11.6 20.4 23.8 -
25.2
## 1362 1428 1433 1456 1772 1815 1866 1876 1879 1929 19
57
## Predicted 3.31 23.65 51.5 45.4 30.20 20.7 17.5 195 -5.48 24.2 -11.
83
## cvpred 2.16 21.72 63.7 44.9 35.66 24.6 21.2 215 -5.73 31.4 -7.
73
## num_deaths 0.00 29.00 53.0 85.0 27.00 0.0 0.0 104 0.00 19.0 0.
00
## CV residual -2.16 7.28 -10.7 40.1 -8.66 -24.6 -21.2 -111 5.73 -12.4 7.
73
## 2010 2062 2106 2117 2122 2140 2195 2289
## Predicted 22.09 55.38 194.6 6.69 26.1 13.2 -5.48 20.1
## cvpred 25.77 58.33 214.6 15.87 33.7 12.6 -5.73 18.4
## num_deaths 33.00 55.00 126.0 0.00 0.0 0.0 0.00 0.0
## CV residual 7.23 -3.33 -88.6 -15.87 -33.7 -12.6 5.73 -18.4
##
## Sum of squares = 219004      Mean square = 4212      n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5311

```

Final First Order Model

##Adjusted R-Squared 0.2344

```

model.firstorder <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West, data = covidcleansed)
summary(model.firstorder)

## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##   influenza + injuryPoisoning + ischemicheart + malignant +
##   obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##   conditionrespiratoryarrest + sepsis + vascular + Northeast +
##   South + West, data = covidcleansed)
##
## Residuals:
##    Min      1Q Median      3Q     Max 
## -147.7  -22.9   -5.8    8.4 1026.0 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  31.57     9.40    3.36  0.00080 ***
## `age45-54`   27.47     3.85    7.14  1.6e-12 ***
## adultrespdistSyn -22.32    12.23   -1.83  0.06818 .  
## allothcondsandcauses 21.04    11.41    1.84  0.06535 .  
## alzheimer    -56.10    10.92   -5.14  3.2e-07 ***
## cardiacArrest -27.85    13.15   -2.12  0.03438 *  
## cardiacarrhythmia -51.02    12.79   -3.99  7.0e-05 *** 
## cerebrovascula -49.77    13.52   -3.68  0.00024 *** 
## chroniclr     -48.24    12.85   -3.75  0.00018 *** 
## covid19        83.82    11.37    7.37  3.0e-13 *** 
## diabetes       -21.64    12.35   -1.75  0.07990 .  
## heartfailure   -49.41    13.52   -3.66  0.00027 *** 
## hypertensive    -25.37    12.71   -2.00  0.04613 *  
## influenza       21.20    11.89    1.78  0.07475 .  
## injuryPoisoning -49.60    12.58   -3.94  8.5e-05 *** 
## ischemicheart  -45.56    13.16   -3.46  0.00055 *** 
## malignant       -50.22    13.35   -3.76  0.00018 *** 
## obesity         -29.45    12.64   -2.33  0.01996 *  
## othercirculatorysystem -44.73    13.32   -3.36  0.00080 *** 
## otherrespsystem -49.51    13.52   -3.66  0.00026 *** 
## renalfailure    -40.64    12.59   -3.23  0.00128 ** 
## conditionrespiratoryarrest -51.85    13.18   -3.93  8.8e-05 *** 
## sepsis          -33.41    12.58   -2.66  0.00800 ** 

```

```

## vascular           -54.43    11.26   -4.83  1.5e-06 ***
## Northeast        14.85     6.01    2.47  0.01358 *
## South            16.16     5.36    3.02  0.00261 **
## West             12.26     5.60    2.19  0.02888 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.3 on 1337 degrees of freedom
## Multiple R-squared:  0.249, Adjusted R-squared:  0.234
## F-statistic: 17.1 on 26 and 1337 DF, p-value: <2e-16

```

Second Order and Interaction terms

Adjusted R-squared: 0.252

MSE 5517

```

library(DAAG)
#covidCleansed$`age45-54SQ`<- covidCleansed$`age45-54`^2
#covidCleansed$covid19SQ <- covidCleansed$covid19^2
model.interactionA <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  adultrespdistSyn*`age45-54` + allothcondsandcauses*`age45-54` +
  alzheimer*`age45-54` + cardiacArrest*`age45-54` + cardiacarrhythmia*`age45-54` +
  cerebrovascula*`age45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` + heartfailure*`age45-54` +
  hypertensive*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-54` +
  malignant*`age45-54` +
  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*`age45-54` +
  renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + sepsis*`age45-54` + vascular*`age45-54` +
  Northeast*`age45-54` +
  South*`age45-54` + West*`age45-54`, data = train)
summary(model.interactionA)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
##     alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##     chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##     influenza + injuryPoisoning + ischemicheart + malignant +
##     obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##     conditionrespiratoryarrest + sepsis + vascular + Northeast +
##     South + West)

```

```

## South + West + adultrespdistSyn * `age45-54` + allothcondsandcauses *
## `age45-54` + alzheimer * `age45-54` + cardiacArrest * `age45-54` +
## cardiacarrhythmia * `age45-54` + cerebrovascula * `age45-54` +
## chroniclr * `age45-54` + covid19 * `age45-54` + diabetes *
## `age45-54` + heartfailure * `age45-54` + hypertensive * `age45-54` +
## influenza * `age45-54` + injuryPoisoning * `age45-54` + ischemicheart
##
## `age45-54` + malignant * `age45-54` + obesity * `age45-54` +
## othercirculatorysystem * `age45-54` + otherrespsystem * `age45-54` +
## renalfailure * `age45-54` + conditionrespiratoryarrest *
## `age45-54` + sepsis * `age45-54` + vascular * `age45-54` +
## Northeast * `age45-54` + South * `age45-54` + West * `age45-54` ,
## data = train)
##
## Residuals:
##      Min    1Q Median    3Q   Max
## -196.1 -17.3   -4.5    4.2 980.9
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   33.385   16.144   2.07  0.03889
## `age45-54`                     29.711   20.831   1.43  0.15409
## adultrespdistSyn            -22.060   21.713  -1.02  0.30987
## allothcondsandcauses          6.768   19.202   0.35  0.72455
## alzheimer                      -38.549   18.336  -2.10  0.03576
## cardiacArrest                  -24.220   21.727  -1.11  0.26523
## cardiacarrhythmia             -37.338   19.928  -1.87  0.06126
## cerebrovascula                 -37.607   20.676  -1.82  0.06921
## chroniclr                      -35.940   21.153  -1.70  0.08961
## covid19                         26.920   19.099   1.41  0.15900
## diabetes                        -12.449   23.282  -0.53  0.59295
## heartfailure                    -36.603   21.746  -1.68  0.09263
## hypertensive                   -22.180   22.027  -1.01  0.31419
## influenza                        6.391   19.762   0.32  0.74648
## injuryPoisoning                -34.484   20.460  -1.69  0.09220
## ischemicheart                  -36.841   20.290  -1.82  0.06970
## malignant                       -36.907   21.204  -1.74  0.08205
## obesity                          -12.594   21.160  -0.60  0.55184
## othercirculatorysystem          -33.268   21.146  -1.57  0.11596
## otherrespsystem                  -35.619   21.751  -1.64  0.10181
## renalfailure                   -24.513   24.299  -1.01  0.31330

```

```

## conditionrespiratoryarrest      -35.842   21.277  -1.68  0.09237
.
## sepsis                         -27.139   20.515  -1.32  0.18616
## vascular                       -38.301   18.429  -2.08  0.03792
*
## Northeast                      4.332    9.785   0.44  0.65805
## South                          9.418    8.622   1.09  0.27492
## West                           4.899    8.627   0.57  0.57025
## `age45-54` :adultrespdistSyn  -9.947   28.382  -0.35  0.72605
## `age45-54` :allothcondsandcauses 0.196   25.911   0.01  0.99395
## `age45-54` :alzheimer          -44.393   24.727  -1.80  0.07289
.
## `age45-54` :cardiacArrest     -26.926   28.704  -0.94  0.34842
## `age45-54` :cardiacarrhythmia -40.090   27.532  -1.46  0.14566
## `age45-54` :cerebrovascula   -40.209   29.229  -1.38  0.16922
## `age45-54` :chroniclr        -34.320   27.965  -1.23  0.22001
## `age45-54` :covid19           86.450    25.367   3.41  0.00068
*** 
## `age45-54` :diabetes          -23.529   28.659  -0.82  0.41184
## `age45-54` :heartfailure     -43.303   33.159  -1.31  0.19186
## `age45-54` :hypertensive     -13.971   28.180  -0.50  0.62017
## `age45-54` :influenza        40.483    26.528   1.53  0.12731
## `age45-54` :injuryPoisoning  -38.883   28.147  -1.38  0.16743
## `age45-54` :ischemicheart   -26.522   28.930  -0.92  0.35947
## `age45-54` :malignant         -36.990   30.278  -1.22  0.22210
## `age45-54` :obesity           -35.597   28.434  -1.25  0.21088
## `age45-54` :othercirculatorysystem -33.620   29.532  -1.14  0.25521
## `age45-54` :otherrespstystem -41.018   30.640  -1.34  0.18095
## `age45-54` :renalfailure      -39.714   29.877  -1.33  0.18406
## `age45-54` :conditionrespiratoryarrest -45.751   29.647  -1.54  0.12308
## `age45-54` :sepsis             -17.866   27.224  -0.66  0.51179
## `age45-54` :vascular           -42.588   25.815  -1.65  0.09930
.
## `age45-54` :Northeast          25.293   13.486   1.88  0.06100
.
## `age45-54` :South              18.222   11.884   1.53  0.12550
## `age45-54` :West               17.885   12.490   1.43  0.15244
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.3 on 1050 degrees of freedom
## Multiple R-squared:  0.287, Adjusted R-squared:  0.252
## F-statistic: 8.27 on 51 and 1050 DF,  p-value: <2e-16

outA <- cv.lm(data = test
                , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +

```



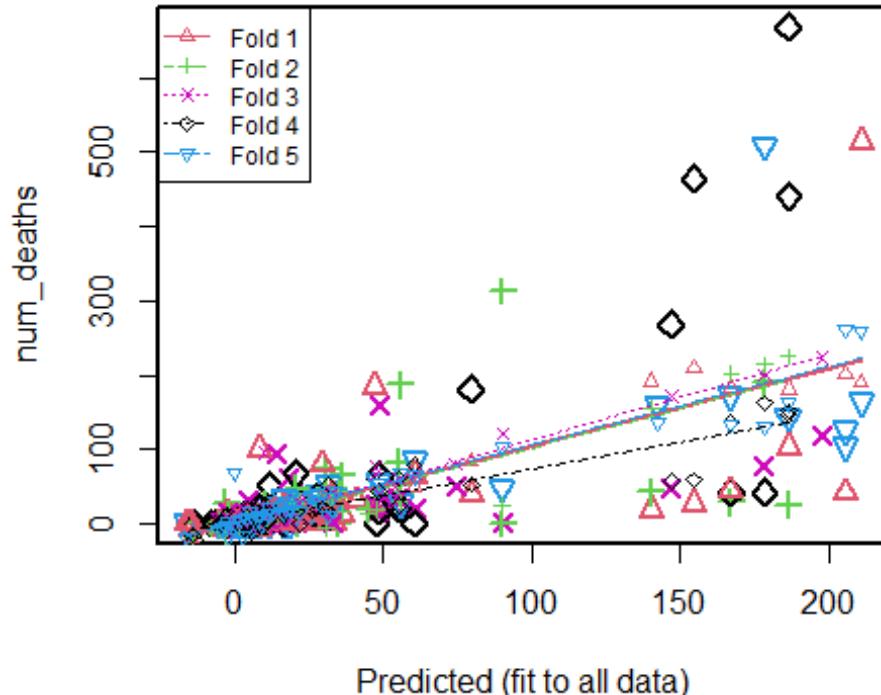
```

## influenza 1 7406 7406 1.52 0.21835
## injuryPoisoning 1 992 992 0.20 0.65182
## ischemicheart 1 500 500 0.10 0.74878
## malignant 1 1759 1759 0.36 0.54805
## obesity 1 5 5 0.00 0.97322
## othercirculatorysystem 1 46 46 0.01 0.92260
## otherrespsystem 1 203 203 0.04 0.83816
## renalfailure 1 4668 4668 0.96 0.32812
## conditionrespiratoryarrest 1 1303 1303 0.27 0.60506
## sepsis 1 176 176 0.04 0.84941
## vascular 1 12311 12311 2.53 0.11293
## Northeast 1 113 113 0.02 0.87876
## South 1 3229 3229 0.66 0.41583
## West 1 2523 2523 0.52 0.47198
## `age45-54` :adultrespdistSyn 1 10 10 0.00 0.96299
## `age45-54` :allothcondsandcauses 1 83232 83232 17.13 5.0e-05
*** 
## `age45-54` :alzheimer 1 1667 1667 0.34 0.55868
## `age45-54` :cardiacArrest 1 3130 3130 0.64 0.42306
## `age45-54` :cardiacarrhythmia 1 583 583 0.12 0.72943
## `age45-54` :cerebrovascula 1 233 233 0.05 0.82677
## `age45-54` :chroniclr 1 486 486 0.10 0.75217
## `age45-54` :covid19 1 5418 5418 1.12 0.29218
## `age45-54` :diabetes 1 733 733 0.15 0.69817
## `age45-54` :heartfailure 1 0 0 0.00 0.99841
## `age45-54` :hypertensive 1 902 902 0.19 0.66697
## `age45-54` :influenza 1 364 364 0.07 0.78458
## `age45-54` :injuryPoisoning 1 178 178 0.04 0.84847
## `age45-54` :ischemicheart 1 1003 1003 0.21 0.65007
## `age45-54` :malignant 1 1042 1042 0.21 0.64381
## `age45-54` :obesity 1 374 374 0.08 0.78172
## `age45-54` :othercirculatorysystem 1 2 2 0.00 0.98298
## `age45-54` :otherrespsystem 1 265 265 0.05 0.81573
## `age45-54` :renalfailure 1 2041 2041 0.42 0.51763
## `age45-54` :conditionrespiratoryarrest 1 291 291 0.06 0.80703
## `age45-54` :sepsis 1 51 51 0.01 0.91855
## `age45-54` :vascular 1 6793 6793 1.40 0.23837
## `age45-54` :Northeast 1 1550 1550 0.32 0.57276
## `age45-54` :South 1 0 0 0.00 0.99424
## `age45-54` :West 1 526 526 0.11 0.74241
## Residuals 210 1020278 4858
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
## :
## 
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function

```

```
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate
```

### Small symbols show cross-validation predicted val



```
##
## fold 1
## Observations in test set: 52
##          21   27   38    79   186   280   359   396   413   418
499
## Predicted -0.644 13.9 8.97 6.74 47.3 79.7 155 5.05 19.03 16.75
33.2
## cvpred 1.155 15.6 4.85 10.44 21.3 83.5 210 11.79 20.73 23.31
42.8
## num_deaths 0.000 0.0 0.00 0.00 185.0 39.0 26 0.00 30.00 21.00
0.0
## CV residual -1.155 -15.6 -4.85 -10.44 163.7 -44.5 -184 -11.79 9.27 -2.31
-42.8
##          509   544   578   609   668   675   699   726   728   810
818
## Predicted 14.2 29.5 -14.17 -3.14 8.42 -2.67 -4.82 186.6 29.5 2.78
206
## cvpred 22.2 37.4 -8.12 -2.15 4.07 -7.25 -8.94 180.5 37.4 8.68
200
## num_deaths 0.0 22.0 0.00 0.00 100.00 13.00 0.00 103.0 10.0 0.00
41
## CV residual -22.2 -15.4 8.12 2.15 95.93 20.25 8.94 -77.5 -27.4 -8.68
-159
##          917   972  1007  1015  1100  1127  1139  1174  1266  1413  1
```

```

464
## Predicted 19.03 9.89 13.898 11.80 36.00 4.33 140 27.64 27.64 25.3 4
0.8
## cvpred 20.73 -9.66 14.951 20.13 20.96 13.00 192 23.98 23.98 27.7 3
7.7
## num_deaths 24.00 0.00 14.000 11.00 11.00 0.00 18 17.00 14.00 0.0 3
1.0
## CV residual 3.27 9.66 -0.951 -9.13 -9.96 -13.00 -174 -6.98 -9.98 -27.7 -
6.7
## 1500 1517 1564 1586 1589 1600 1689 1892 1909 1914
## Predicted -5.11 -0.483 21.26 60.5 -0.0703 211 33.0 -14.6 4.33 -16.4
## cvpred -10.90 0.770 12.92 50.1 -5.0141 191 43.8 -19.9 13.00 -11.2
## num_deaths 0.00 0.000 14.00 64.0 0.0000 514 10.0 0.0 0.00 0.0
## CV residual 10.90 -0.770 1.08 13.9 5.0141 323 -33.8 19.9 -13.00 11.2
## 1965 1994 2071 2138 2174 2236 2275 2304 2313
## Predicted 32.97 29.7 8.34 10.27 14.04 167 12.8 21.80 0.722
## cvpred 43.76 26.9 13.48 2.13 7.71 184 17.9 7.29 2.100
## num_deaths 39.00 79.0 0.00 0.00 15.00 44 0.0 0.00 0.000
## CV residual -4.76 52.1 -13.48 -2.13 7.29 -140 -17.9 -7.29 -2.100
##
## Sum of squares = 268909 Mean square = 5171 n = 52
##
## fold 2
## Observations in test set: 53
## 39 125 156 178 198 266 310 352 355 371
## Predicted 20.9 6.74 8.35 178.5 90.3 47.88 8.0 2.78 11.8 8.34
## cvpred 26.1 11.70 -4.10 214.1 24.0 19.19 15.5 2.82 19.5 12.30
## num_deaths 0.0 0.00 11.00 190.0 313.0 22.00 0.0 0.00 0.0 0.00
## CV residual -26.1 -11.70 15.10 -24.1 289.0 2.81 -15.5 -2.82 -19.5 -12.30
## 397 412 457 458 492 503 517 571 586 635 6
56
## Predicted 6.74 55.2 28.4 55.22 55.8 16.3 -3.14 -2.67 167 30.78 -2.
98
## cvpred 11.70 51.3 31.7 51.34 29.8 18.6 -2.16 1.80 202 -5.26 -16.
00
## num_deaths 0.00 82.0 15.0 53.00 42.0 0.0 0.00 0.00 29 74.00 28.
00
## CV residual -11.70 30.7 -16.7 1.66 12.2 -18.6 2.16 -1.80 -173 79.26 44.
00
## 736 876 894 900 977 982 1003 1028 1034 1039 104
1
## Predicted -3.19 22.0 11.15 10.3 1.22 34.9 30.78 10.5 36.1 -7.72 -1.7
1
## cvpred -7.76 31.4 18.91 16.8 -3.27 49.0 -5.26 13.8 15.9 -6.59 -7.8
8
## num_deaths 0.00 10.0 11.00 0.0 0.00 0.0 46.00 25.0 65.0 0.00 12.0
0
## CV residual 7.76 -21.4 -7.91 -16.8 3.27 -49.0 51.26 11.2 49.1 6.59 19.8
8
## 1084 1204 1222 1239 1288 1321 1511 1559 1567 1645 1675

```

```

1690
## Predicted -14.2 90.3 -2.83 19.0 -3.19 18.51 4.46 20.7 18.59 140 11.8
186
## cvpred -18.4 24.0 -6.53 21.8 -7.76 20.85 7.69 17.5 17.55 156 20.1
225
## num_deaths 0.0 0.0 0.00 14.0 0.00 14.00 0.00 47.0 12.00 43 0.0
25
## CV residual 18.4 -24.0 6.53 -7.8 7.76 -6.85 -7.69 29.5 -5.55 -113 -20.1
-200
## 1734 1773 1873 1983 2008 2035 2061 2097 2211
## Predicted 8.97 -0.0703 32.97 14.50 55.8 -3.14 45.24 6.74 -4.82
## cvpred 10.20 -4.0398 39.15 15.67 29.8 -2.16 13.04 11.70 -9.15
## num_deaths 0.00 0.0000 42.00 17.00 187.0 0.00 20.00 0.00 0.00
## CV residual -10.20 4.0398 2.85 1.33 157.2 2.16 6.96 -11.70 9.15
##
## Sum of squares = 215029 Mean square = 4057 n = 53
##
## fold 3
## Observations in test set: 53
## 16 190 209 211 274 307 325 428 454
487
## Predicted 90.3 8.15 -0.241 13.86 53.9 -0.0703 19.0 16.236 48.7
7.6
## cvpred 120.3 21.44 -10.574 -4.04 49.3 7.9720 19.1 17.341 38.6 1
0.6
## num_deaths 0.0 14.00 13.000 93.00 23.0 0.0000 0.0 17.000 158.0
0.0
## CV residual -120.3 -7.44 23.574 97.04 -26.3 -7.9720 -19.1 -0.341 119.4 -1
0.6
## 491 527 597 677 774 793 868 904 907 909
967
## Predicted 12.75 13.86 -0.483 18.51 48.7 8.89 10.0 61.0 25.3 147 0
.67
## cvpred 16.41 -4.04 -4.409 5.75 38.6 15.31 94.7 67.3 24.3 171 7
.53
## num_deaths 12.00 0.00 0.000 61.00 18.0 0.00 10.0 21.0 40.0 48 0
.00
## CV residual -4.41 4.04 4.409 55.25 -20.6 -15.31 -84.7 -46.3 15.7 -123 -7
.53
## 994 1009 1014 1071 1213 1225 1327 1334 1375 1
391
## Predicted 8.0 4.57 -2.47 -3.07 -2.93e-14 0.722 2.03 -3.19 20.7 1
.22
## cvpred 10.8 -2.02 -7.32 -2.67 -6.69e+01 -2.078 2.41 -5.63 30.6 12
.72
## num_deaths 0.0 32.00 15.00 0.00 0.00e+00 0.000 0.00 0.00 0.00 0.0
.00
## CV residual -10.8 34.02 22.32 2.67 6.69e+01 2.078 -2.41 5.63 -30.6 -12
.72
## 1438 1472 1523 1540 1554 1760 1805 1818 1836 1838

```

```

1936
## Predicted    21.5   8.15   2.466   47.4   198   21.5   1.22   32.9   55.2   74.4
10.44
## cvpred      28.0   21.44   0.591   76.1   223   28.0   12.72   41.3   59.5   79.6
-6.85
## num_deaths  13.0   0.00   0.000   15.0   119   11.0   0.00   0.0   33.0   49.0
13.00
## CV residual -15.0  -21.44  -0.591  -61.1  -104  -17.0  -12.72  -41.3  -26.5  -30.6
19.85
##          1960     2041   2050   2052   2125   2152   2190     2225   2248
2252
## Predicted    2.775  -2.93e-14  -2.83  -5.11   3.72   47.3   179  4.33e+00  16.8
4.66
## cvpred       0.177  -6.69e+01   8.10   7.03   15.90   66.3   199  1.39e-13  12.9
4.78
## num_deaths   0.000  0.00e+00   0.00   0.00   0.00   25.0   78  1.30e+01   0.0
0.00
## CV residual -0.177  6.69e+01  -8.10  -7.03  -15.90  -41.3  -121  1.30e+01  -12.9
-4.78
##          2255
## Predicted    9.64
## cvpred       21.17
## num_deaths   0.00
## CV residual -21.17
##
## Sum of squares = 117367      Mean square = 2214      n = 53
##
## fold 4
## Observations in test set: 52
##          7   94   126   159   183   217   313   320   382   385
434
## Predicted    7.0   55.2   2.78  -1.02   4.458  -5.292   4.12  48.7  16.24  11.02  2
1.922
## cvpred      18.1   61.4  -6.29   3.20  -0.448   0.213   2.25  43.5   9.46   2.73  -
0.575
## num_deaths  0.0   23.0   0.00  10.00  15.000   0.000   0.00  63.0   0.00   0.00  2
5.000
## CV residual -18.1  -38.4   6.29   6.80  15.448  -0.213  -2.25  19.5  -9.46  -2.73  2
5.575
##          550   628   630   647   680   761   808   881   936   944   96
2
## Predicted   55.22  47.9   8.97  11.80  187  -6.86   5.05  31.5  29.68  27.6   60.
4
## cvpred      24.79  71.3   17.92  7.22  151   1.67  -5.04  53.0  20.33  20.2   78.
9
## num_deaths  21.00  0.0   0.00  52.00  440   0.00  0.00  10.0  17.00  33.0   0.
0
## CV residual -3.79  -71.3  -17.92  44.78  289  -1.67   5.04  -43.0  -3.33  12.8  -78.
9
##          963   965   992  1083  1136  1138  1171  1197  1221  1308

```

```

1315
## Predicted    11.4  3.88  10.3 -6.86 -2.37 167.2 25.89 -3.692 -4.43 -13.3 -
7.722
## cvpred     16.1 -2.24  12.5  1.67 -8.10 136.4 18.35  0.994  1.08 -23.0
0.811
## num_deaths   0.0  0.00   0.0  0.00   0.00  42.0 12.00  0.000  0.00   0.0
0.000
## CV residual -16.1  2.24 -12.5 -1.67  8.10 -94.4 -6.35 -0.994 -1.08  23.0 -
0.811
##          1421 1461 1490 1506 1513  1544 1562 1626 1636 1648 1775
## Predicted   20.7 147.1   14  179 16.3 -2.83 79.7 10.46 -14.2 48.7 -0.93
## cvpred      15.2 57.2    15 162 17.2 10.54 52.8 9.89 -15.5 43.5 -1.83
## num_deaths  67.0 268.0    10  42 15.0  0.00 180.0 16.00   0.0 17.0  0.00
## CV residual 51.8 210.8   -5 -120 -2.2 -10.54 127.2 6.11 15.5 -26.5 1.83
##          1776 1814 1860 1883 1979 2012 2013 2019
## Predicted   8.0 11.15   5.92  4.57 8.341 186 154.8 16.3
## cvpred      11.3 4.98 -12.56 18.70 0.525 147 58.8 17.2
## num_deaths   0.0  0.00   0.00  0.00 22.000 668 464.0  0.0
## CV residual -11.3 -4.98 12.56 -18.70 21.475 521 405.2 -17.2
##
## Sum of squares = 628555      Mean square = 12088      n = 52
##
## fold 5
## Observations in test set: 52
##          5     12     26     31    123    152    176    179    204    224     3
06
## Predicted   16.3   2.56  -1.96 -4.43   7.60   90.3   8.970  143 34.00 179 32
.9
## cvpred      14.9 -17.39 -18.98 -5.50   8.95 104.3 -0.406 137 43.57 132 36
.2
## num_deaths   0.0   0.00   0.00   0.00   0.00  48.0 15.000 160 34.00 509 10
.0
## CV residual -14.9  17.39 18.98  5.50  -8.95 -56.3 15.406  23 -9.57 377 -26
.2
##          411    486    533    537    757    840    849    858    871    910
946
## Predicted   28.4  27.6  -5.29 -0.241   11.4 16.24   25.9   55.8 11.36 211.1
5.05
## cvpred      29.4 39.1  -8.22 -3.149   22.6 24.17   37.1   67.1 12.34 258.6  1
6.93
## num_deaths  40.0  25.0   0.00  0.000    0.0 17.00   18.0   28.0 22.00 164.0
0.00
## CV residual 10.6 -14.1   8.22  3.149 -22.6 -7.17 -19.1 -39.1  9.66 -94.6 -1
6.93
##          949    954    985   1046   1090   1091   1207   1216   1287     129
3
## Predicted   12.755 186.4   4.12 167.2 -2.37 18.51 -3.14  -1.96 -2.91  0.035
8
## cvpred     13.484 161.9   7.59 133.2  1.22 18.27 -5.52 -18.98 -2.75  1.212
5

```

```

## num_deaths 14.000 143.0 0.00 172.0 0.00 20.00 0.00 0.00 0.00 0.00 0.000
0
## CV residual 0.516 -18.9 -7.59 38.8 -1.22 1.73 5.52 18.98 2.75 -1.212
5
## 1312 1362 1428 1433 1456 1772 1815 1866 1876 1879 192
9
## Predicted 8.42 -16.4 22.0 31.5 61.0 32.9 18.2 5.05 206 2.03 19.03
0
## cvpred 10.38 -18.9 16.5 10.0 64.2 36.2 30.7 16.93 262 4.96 18.73
5
## num_deaths 0.00 0.0 29.0 53.0 85.0 27.0 0.0 0.00 104 0.00 19.00
0
## CV residual -10.38 18.9 12.5 43.0 20.8 -9.2 -30.7 -16.93 -158 -4.96 0.26
5
## 1957 2010 2062 2106 2117 2122 2140 2195 2289
## Predicted 7.60 16.85 48.677 206 18.6 0.0 8.0 2.03 7.00
## cvpred 8.95 29.98 54.107 262 24.8 68.9 6.9 4.96 4.28
## num_deaths 0.00 33.00 55.000 126 0.0 0.0 0.0 0.00 0.00
## CV residual -8.95 3.02 0.893 -136 -24.8 -68.9 -6.9 -4.96 -4.28
##
## Sum of squares = 215695 Mean square = 4148 n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5517

```

##Prune out *age45-54:adultrespdistSyn* ## Adjusted R-squared: 0.253 ## MSE 5442

```

model.interactionB <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothc
ondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` +
  alzheimer*`age45-54` + cardiacArrest*`age45-54` + cardiacarrhythmia*`age4
5-54` + cerebrovascula*`age45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` + heartfa
ilure*`age45-54` + hypertensive*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-
54` + malignant*`age45-54` +
  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*`a
ge45-54` + renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + sepsis*`age45-54` + vascular*`age
45-54` + Northeast*`age45-54` +
  South*`age45-54` + West*`age45-54` , data = train)
summary(model.interactionB)

```

```

## 
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##     alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##     chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##     influenza + injuryPoisoning + ischemicheart + malignant +
##     obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##     conditionrespiratoryarrest + sepsis + vascular + Northeast +
##     South + West + allothcondsandcauses * `age45-54` + alzheimer *
##     `age45-54` + cardiacArrest * `age45-54` + cardiacarrhythmia *
##     `age45-54` + cerebrovascula * `age45-54` + chroniclr * `age45-54` +
##     covid19 * `age45-54` + diabetes * `age45-54` + heartfailure *
##     `age45-54` + hypertensive * `age45-54` + influenza * `age45-54` +
##     injuryPoisoning * `age45-54` + ischemicheart * `age45-54` +
##     malignant * `age45-54` + obesity * `age45-54` + othercirculatorysystem
##
##     `age45-54` + otherrespsystem * `age45-54` + renalfailure *
##     `age45-54` + conditionrespiratoryarrest * `age45-54` + sepsis *
##     `age45-54` + vascular * `age45-54` + Northeast * `age45-54` +
##     South * `age45-54` + West * `age45-54`, data = train)
## 
## Residuals:
##    Min      1Q Median      3Q     Max
## -196.0   -17.1   -4.5    4.2   980.9
## 
## Coefficients:
## (Intercept)            36.09   14.17   2.55   0.011
## `age45-54`              25.25   16.47   1.53   0.126
## adultrespdistSyn       -27.88   13.98  -1.99   0.046
## allothcondsandcauses   4.07    17.58   0.23   0.817
## alzheimer             -41.25   16.63  -2.48   0.013
## cardiacArrest          -26.91   20.32  -1.32   0.186
## cardiacarrhythmia      -40.03   18.38  -2.18   0.030
## cerebrovascula         -40.30   19.18  -2.10   0.036
## chroniclr              -38.64   19.69  -1.96   0.050
## .
## covid19                 24.22   17.47   1.39   0.166
## diabetes                -15.16   21.95  -0.69   0.490
## heartfailure            -39.30   20.33  -1.93   0.054
## .
## hypertensive            -24.88   20.63  -1.21   0.228
## influenza                3.69    18.19   0.20   0.839
## injuryPoisoning         -37.18   18.95  -1.96   0.050

```

|                                                                  |        |       |       |         |  |
|------------------------------------------------------------------|--------|-------|-------|---------|--|
| *                                                                |        |       |       |         |  |
| ## ischemicheart                                                 | -39.53 | 18.77 | -2.11 | 0.035   |  |
| *                                                                |        |       |       |         |  |
| ## malignant                                                     | -39.60 | 19.75 | -2.00 | 0.045   |  |
| *                                                                |        |       |       |         |  |
| ## obesity                                                       | -15.29 | 19.70 | -0.78 | 0.438   |  |
| ## othercirculatorysystem                                        | -35.96 | 19.69 | -1.83 | 0.068   |  |
| .                                                                |        |       |       |         |  |
| ## otherrespsystem                                               | -38.31 | 20.34 | -1.88 | 0.060   |  |
| .                                                                |        |       |       |         |  |
| ## renalfailure                                                  | -27.21 | 23.04 | -1.18 | 0.238   |  |
| ## conditionrespiratoryarrest                                    | -38.54 | 19.82 | -1.94 | 0.052   |  |
| .                                                                |        |       |       |         |  |
| ## sepsis                                                        | -29.84 | 19.00 | -1.57 | 0.117   |  |
| ## vascular                                                      | -41.00 | 16.73 | -2.45 | 0.014   |  |
| *                                                                |        |       |       |         |  |
| ## Northeast                                                     | 4.33   | 9.78  | 0.44  | 0.658   |  |
| ## South                                                         | 9.43   | 8.62  | 1.09  | 0.274   |  |
| ## West                                                          | 4.85   | 8.62  | 0.56  | 0.574   |  |
| ## `age45-54` :allothcondsandcauses                              | 4.65   | 22.57 | 0.21  | 0.837   |  |
| ## `age45-54` :alzheimer                                         | -39.94 | 21.21 | -1.88 | 0.060   |  |
| .                                                                |        |       |       |         |  |
| ## `age45-54` :cardiacArrest                                     | -22.48 | 25.74 | -0.87 | 0.383   |  |
| ## `age45-54` :cardiacarrhythmia                                 | -35.64 | 24.42 | -1.46 | 0.145   |  |
| ## `age45-54` :cerebrovascula                                    | -35.75 | 26.31 | -1.36 | 0.174   |  |
| ## `age45-54` :chroniclr                                         | -29.87 | 24.90 | -1.20 | 0.231   |  |
| ## `age45-54` :covid19                                           | 90.90  | 21.95 | 4.14  | 3.7e-05 |  |
| ***                                                              |        |       |       |         |  |
| ## `age45-54` :diabetes                                          | -19.07 | 25.67 | -0.74 | 0.458   |  |
| ## `age45-54` :heartfailure                                      | -38.87 | 30.63 | -1.27 | 0.205   |  |
| ## `age45-54` :hypertensive                                      | -9.52  | 25.14 | -0.38 | 0.705   |  |
| ## `age45-54` :influenza                                         | 44.93  | 23.28 | 1.93  | 0.054   |  |
| .                                                                |        |       |       |         |  |
| ## `age45-54` :injuryPoisoning                                   | -34.42 | 25.09 | -1.37 | 0.170   |  |
| ## `age45-54` :ischemicheart                                     | -22.07 | 25.99 | -0.85 | 0.396   |  |
| ## `age45-54` :malignant                                         | -32.53 | 27.46 | -1.18 | 0.236   |  |
| ## `age45-54` :obesity                                           | -31.15 | 25.43 | -1.22 | 0.221   |  |
| ## `age45-54` :othercirculatorysystem                            | -29.17 | 26.65 | -1.09 | 0.274   |  |
| ## `age45-54` :otherrespsystem                                   | -36.57 | 27.88 | -1.31 | 0.190   |  |
| ## `age45-54` :renalfailure                                      | -35.27 | 27.04 | -1.30 | 0.192   |  |
| ## `age45-54` :conditionrespiratoryarrest                        | -41.29 | 26.76 | -1.54 | 0.123   |  |
| ## `age45-54` :sepsis                                            | -13.41 | 24.06 | -0.56 | 0.577   |  |
| ## `age45-54` :vascular                                          | -38.14 | 22.47 | -1.70 | 0.090   |  |
| .                                                                |        |       |       |         |  |
| ## `age45-54` :Northeast                                         | 25.23  | 13.48 | 1.87  | 0.061   |  |
| .                                                                |        |       |       |         |  |
| ## `age45-54` :South                                             | 18.24  | 11.88 | 1.54  | 0.125   |  |
| ## `age45-54` :West                                              | 17.95  | 12.48 | 1.44  | 0.151   |  |
| ## ---                                                           |        |       |       |         |  |
| ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 |        |       |       |         |  |

```

## 
## Residual standard error: 69.3 on 1051 degrees of freedom
## Multiple R-squared:  0.287, Adjusted R-squared:  0.253
## F-statistic: 8.44 on 50 and 1051 DF,  p-value: <2e-16

outB <- cv.lm(data = test
               , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` +
  alzheimer*`age45-54` + cardiacArrest*`age45-54` + cardiacarrhythmia*`age4
5-54` + cerebrovascula*`age45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` + heartfa
ilure*`age45-54` + hypertensive*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-
54` + malignant*`age45-54` +
  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*
`age45-54` + renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + sepsis*`age45-54` + vascular*`age
45-54` + Northeast*`age45-54` +
  South*`age45-54` + West*`age45-54`)
               , plotit = "Observed", m=5)

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Analysis of Variance Table
## 
## Response: num_deaths
## 
## `age45-54`                               Df  Sum Sq Mean Sq F value Pr(>F)
## 1      55638    55638   11.49 0.00083
## ***

## adultrespdistSyn                         1      308     308    0.06 0.80108
## allothcondsandcauses                     1 124916 124916   25.80 8.3e-07
## ***

```

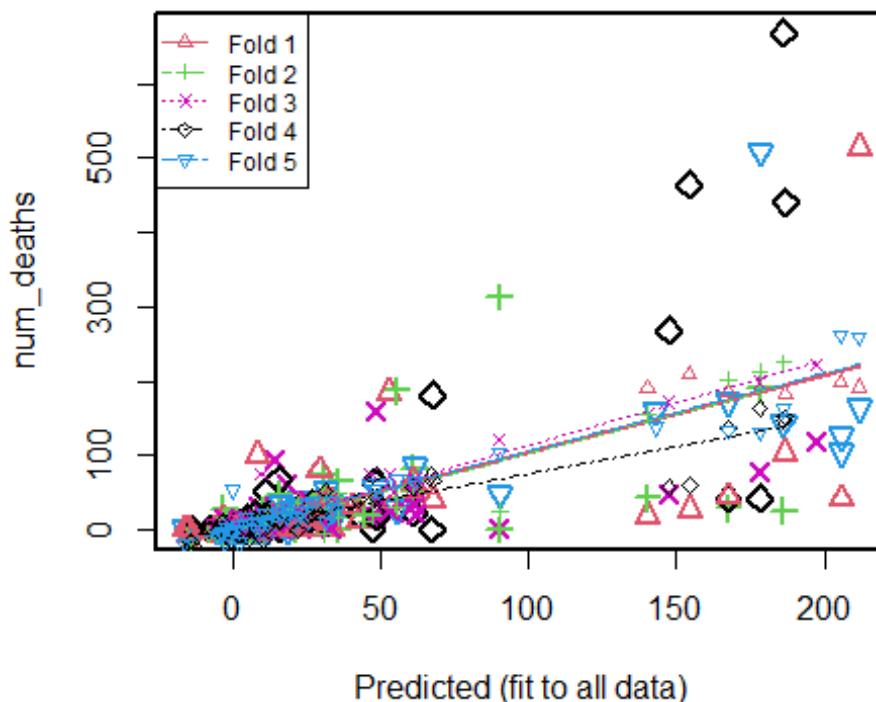
|                                           |   |        |        |       |         |
|-------------------------------------------|---|--------|--------|-------|---------|
| ## alzheimer                              | 1 | 17801  | 17801  | 3.68  | 0.05654 |
| .                                         |   |        |        |       |         |
| ## cardiacArrest                          | 1 | 9110   | 9110   | 1.88  | 0.17162 |
| ## cardiacarrhythmia                      | 1 | 1509   | 1509   | 0.31  | 0.57722 |
| ## cerebrovascula                         | 1 | 512    | 512    | 0.11  | 0.74535 |
| ## chroniclr                              | 1 | 1137   | 1137   | 0.23  | 0.62846 |
| ## covid19                                | 1 | 355946 | 355946 | 73.51 | 2.1e-15 |
| ***                                       |   |        |        |       |         |
| ## diabetes                               | 1 | 3533   | 3533   | 0.73  | 0.39399 |
| ## heartfailure                           | 1 | 499    | 499    | 0.10  | 0.74843 |
| ## hypertensive                           | 1 | 196    | 196    | 0.04  | 0.84063 |
| ## influenza                              | 1 | 7406   | 7406   | 1.53  | 0.21756 |
| ## injuryPoisoning                        | 1 | 992    | 992    | 0.20  | 0.65126 |
| ## ischemicheart                          | 1 | 500    | 500    | 0.10  | 0.74837 |
| ## malignant                              | 1 | 1759   | 1759   | 0.36  | 0.54738 |
| ## obesity                                | 1 | 5      | 5      | 0.00  | 0.97317 |
| ## othercirculatorysystem                 | 1 | 46     | 46     | 0.01  | 0.92247 |
| ## otherrespsystem                        | 1 | 203    | 203    | 0.04  | 0.83789 |
| ## renalfailure                           | 1 | 4668   | 4668   | 0.96  | 0.32730 |
| ## conditionrespiratoryarrest             | 1 | 1303   | 1303   | 0.27  | 0.60445 |
| ## sepsis                                 | 1 | 176    | 176    | 0.04  | 0.84916 |
| ## vascular                               | 1 | 12311  | 12311  | 2.54  | 0.11232 |
| ## Northeast                              | 1 | 113    | 113    | 0.02  | 0.87856 |
| ## South                                  | 1 | 3229   | 3229   | 0.67  | 0.41504 |
| ## West                                   | 1 | 2523   | 2523   | 0.52  | 0.47123 |
| ## `age45-54` :allothcondsandcauses       | 1 | 82454  | 82454  | 17.03 | 5.3e-05 |
| ***                                       |   |        |        |       |         |
| ## `age45-54` :alzheimer                  | 1 | 1892   | 1892   | 0.39  | 0.53263 |
| ## `age45-54` :cardiacArrest              | 1 | 2956   | 2956   | 0.61  | 0.43549 |
| ## `age45-54` :cardiacarrhythmia          | 1 | 618    | 618    | 0.13  | 0.72117 |
| ## `age45-54` :cerebrovascula             | 1 | 200    | 200    | 0.04  | 0.83931 |
| ## `age45-54` :chroniclr                  | 1 | 526    | 526    | 0.11  | 0.74203 |
| ## `age45-54` :covid19                    | 1 | 4981   | 4981   | 1.03  | 0.31163 |
| ## `age45-54` :diabetes                   | 1 | 600    | 600    | 0.12  | 0.72526 |
| ## `age45-54` :heartfailure               | 1 | 21     | 21     | 0.00  | 0.94799 |
| ## `age45-54` :hypertensive               | 1 | 1086   | 1086   | 0.22  | 0.63627 |
| ## `age45-54` :influenza                  | 1 | 518    | 518    | 0.11  | 0.74401 |
| ## `age45-54` :injuryPoisoning            | 1 | 289    | 289    | 0.06  | 0.80738 |
| ## `age45-54` :ischemicheart              | 1 | 1170   | 1170   | 0.24  | 0.62355 |
| ## `age45-54` :malignant                  | 1 | 1248   | 1248   | 0.26  | 0.61215 |
| ## `age45-54` :obesity                    | 1 | 489    | 489    | 0.10  | 0.75101 |
| ## `age45-54` :othercirculatorysystem     | 1 | 13     | 13     | 0.00  | 0.95796 |
| ## `age45-54` :otherrespsystem            | 1 | 364    | 364    | 0.08  | 0.78429 |
| ## `age45-54` :renalfailure               | 1 | 1639   | 1639   | 0.34  | 0.56134 |
| ## `age45-54` :conditionrespiratoryarrest | 1 | 523    | 523    | 0.11  | 0.74265 |
| ## `age45-54` :sepsis                     | 1 | 125    | 125    | 0.03  | 0.87238 |
| ## `age45-54` :vascular                   | 1 | 5683   | 5683   | 1.17  | 0.27990 |
| ## `age45-54` :Northeast                  | 1 | 1685   | 1685   | 0.35  | 0.55592 |
| ## `age45-54` :South                      | 1 | 3      | 3      | 0.00  | 0.98037 |
| ## `age45-54` :West                       | 1 | 389    | 389    | 0.08  | 0.77725 |

```

## Residuals                               211 1021681     4842
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
## :
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

## Small symbols show cross-validation predicted values



```

##
## fold 1
## Observations in test set: 52
##          21    27    38     79   186   280   359    396    413    41
## Predicted -0.44  14.2  8.72   6.43  53.1  67.8  154   4.74 23.1779 16.23
## cvpred     2.25  16.4  4.10  10.17  35.8  54.8  209  10.78 30.0382 21.74
## num_deaths  0.00   0.0   0.00    0.00 185.0  39.0   26   0.00 30.0000 21.00
## CV residual -2.25 -16.4 -4.10 -10.17 149.2 -15.8 -183 -10.78 -0.0382 -0.74
##          499   509   544   578   609   668   675   699   726   728
## 810

```

```

## Predicted    33.7 14.5 29.8 -13.96 -3.01    8.48 -2.86 -4.74 186.8 29.8
2.38
## cvpred     44.5 23.0 39.8 -6.37 -1.86    5.50 -8.66 -9.28 182.3 39.8
7.56
## num_deaths  0.0  0.0 22.0  0.00  0.00 100.00 13.00  0.00 103.0 10.0
0.00
## CV residual -44.5 -23.0 -17.8   6.37  1.86  94.50 21.66  9.28 -79.3 -29.8
-7.56
##           818  917  972 1007 1015 1100 1127 1139 1174 1266 14
13
## Predicted   205 23.18 10.41  8.68 11.51  42.5  4.33 140 27.18 27.18 25
.7
## cvpred      199 30.04 -7.91  3.75 19.03  37.9 13.00 191 22.65 22.65 28
.6
## num_deaths   41 24.00  0.00 14.00 11.00  11.0  0.00 18 17.00 14.00  0
.0
## CV residual -158 -6.04  7.91 10.25 -8.03 -26.9 -13.00 -173 -5.65 -8.65 -28
.6
##           1464 1500 1517 1564 1586 1589 1600 1689 1892 1909
1914
## Predicted   40.36 -5.73 -0.33 21.95 61.1  0.219 212 32.6 -14.6  4.33 -
16.32
## cvpred      37.76 -11.66  1.50 12.87 50.3 -4.381 192 43.7 -17.1 13.00
-9.59
## num_deaths   31.00  0.00  0.00 14.00 64.0  0.000 514 10.0  0.0  0.00
0.00
## CV residual -6.76 11.66 -1.50  1.13 13.7  4.381 322 -33.7 17.1 -13.00
9.59
##           1965 1994 2071 2138 2174 2236 2275 2304 2313
## Predicted   32.64 29.4  8.04 11.01 13.96 168 12.4 21.26 1.21
## cvpred      43.73 25.0 12.92  2.87  7.89 185 17.8 5.83 3.20
## num_deaths  39.00 79.0  0.00  0.00 15.00  44  0.0  0.00  0.00
## CV residual -4.73 54.0 -12.92 -2.87  7.11 -141 -17.8 -5.83 -3.20
##
## Sum of squares = 262828      Mean square = 5054      n = 52
##
## fold 2
## Observations in test set: 53
##           39 125 156 178 198 266 310 352 355 371
## Predicted   21.4 6.43 7.78 178.2 90.3 47.41 8.65 2.38 11.4 8.04
## cvpred      26.7 11.05 -5.14 213.5 24.0 18.24 16.20 2.66 18.7 11.84
## num_deaths  0.0  0.00 11.00 190.0 313.0 22.00 0.00 0.00 0.0  0.00
## CV residual -26.7 -11.05 16.14 -23.5 289.0 3.76 -16.20 -2.66 -18.7 -11.84
##           397 412 457 458 492 503 517 571 586 635 656
## Predicted   6.43 61.2 22.89 61.22 55.5 11.6 -3.01 -2.86 168 30.7 -2.8
## cvpred      11.05 58.8 24.61 58.79 29.6 12.6 -1.88 1.35 203 -5.8 -16.1
## num_deaths  0.00 82.0 15.00 53.00 42.0 0.0  0.00 0.00 29 74.0 28.0
## CV residual -11.05 23.2 -9.61 -5.79 12.4 -12.6 1.88 -1.35 -174 79.8 44.1
##           736 876 894 900 977 982 1003 1028 1034 1039 1041
## Predicted   -3.02 22.5 11.83 11.0 1.27 35.6 30.7 10.7 36.1 -7.78 -1.76

```

```

## cvpred      -7.44  32.0 19.82  17.6 -3.19  49.9 -5.8 14.1 16.1 -6.79 -7.82
## num_deaths  0.00  10.0 11.00   0.0  0.00   0.0 46.0 25.0 65.0  0.00 12.00
## CV residual 7.44 -22.0 -8.82 -17.6  3.19 -49.9 51.8 10.9 48.9  6.79 19.82
##               1084 1204 1222 1239 1288 1321 1511 1559 1567 1645 167
5
## Predicted   -14.0  90.3 -3.37  23.2 -3.02 18.43  4.78 15.9 18.76  140 11.
4
## cvpred      -18.3  24.0 -7.31  25.8 -7.44 20.89  8.10 11.3 17.67  156 19.
2
## num_deaths   0.0   0.0  0.00  14.0  0.00 14.00  0.00 47.0 12.00   43  0.
0
## CV residual  18.3 -24.0  7.31 -11.8  7.44 -6.89 -8.10 35.7 -5.67 -113 -19.
2
##               1690 1734 1773 1873 1983 2008 2035 2061 2097 2211
## Predicted    186  8.72  0.219 32.64 14.15 55.5 -3.01 44.92  6.43 -4.74
## cvpred       225  9.77 -3.737 38.72 14.93 29.6 -1.88 12.03 11.05 -8.92
## num_deaths   25  0.00  0.000 42.00 17.00 187.0  0.00 20.00  0.00  0.00
## CV residual  -200 -9.77  3.737  3.28  2.07 157.4  1.88  7.97 -11.05  8.92
##
## Sum of squares = 215297      Mean square = 4062      n = 53
##
## fold 3
## Observations in test set: 53
##               16   190   209   211   274   307   325   428   454
487
## Predicted   90.3  7.56  0.109 14.17  54.7  0.219  23.2 15.8967 48.5  7
.18
## cvpred     120.3 20.15 -9.980 -3.51  51.0  7.981  23.9 17.0238 38.2 10
.13
## num_deaths  0.0 14.00 13.000 93.00  23.0  0.000   0.0 17.0000 158.0  0
.00
## CV residual -120.3 -6.15 22.980 96.51 -28.0 -7.981 -23.9 -0.0238 119.8 -10
.13
##               491   527   597   677   774   793   868   904   907   909   9
67
## Predicted   12.45 14.17 -0.33 18.43  48.5  8.23  10.0 61.8 25.7 147  1.
08
## cvpred     16.09 -3.51 -4.11  5.86  38.2  14.52  74.6 69.0 24.5 171  7.
76
## num_deaths 12.00  0.00  0.00 61.00  18.0  0.00  10.0 21.0 40.0  48  0.
00
## CV residual -4.09  3.51  4.11 55.14 -20.2 -14.52 -64.6 -48.0 15.5 -123 -7.
76
##               994 1009 1014 1071           1213 1225 1327 1334 1375  1
391
## Predicted   8.65  8.97 -2.47 -3.40 -8.79e-14  1.21  1.81 -3.02 15.9  1
.27
## cvpred     11.97  3.10 -7.60 -3.14 -4.45e+01 -1.52  2.18 -5.95 25.4 12
.37
## num_deaths  0.00 32.00 15.00  0.00  0.00e+00  0.00  0.00  0.00  0.00  0.0  0

```

```

.00
## CV residual -11.97 28.90 22.60 3.14 4.45e+01 1.52 -2.18 5.95 -25.4 -12
.37
##          1438   1472   1523   1540   1554   1760   1805   1818   1836   1838  1
936
## Predicted 22.2    7.56   2.91   46.7   197    22.2   1.27   33.4   61.2   61.5  1
0.6
## cvpred    29.8   20.15   1.26   74.7   222    29.8   12.37  42.6   66.5   62.7  -
7.0
## num_deaths 13.0    0.00   0.00   15.0   119    11.0   0.00   0.0    33.0   49.0  1
3.0
## CV residual -16.8 -20.15 -1.26 -59.7 -103   -18.8 -12.37 -42.6 -33.5 -13.7  2
0.0
##          1960     2041   2050   2052   2125   2152   2190   2225   2248
2252
## Predicted 2.379 -8.79e-14 -3.37 -5.73   3.84   53.1   178   4.33e+00  16.2
4.14
## cvpred    -0.787 -4.45e+01  7.42   6.22   15.47  73.5   199   8.88e-14  11.5
3.17
## num_deaths 0.000  0.00e+00  0.00   0.00   0.00   25.0   78   1.30e+01  0.0
0.00
## CV residual 0.787  4.45e+01 -7.42 -6.22 -15.47 -48.5 -121  1.30e+01 -11.5
-3.17
##          2255
## Predicted 9.47
## cvpred    20.75
## num_deaths 0.00
## CV residual -20.75
##
## Sum of squares = 109121      Mean square = 2059      n = 53
##
## fold 4
## Observations in test set: 52
##          7   94   126   159   183   217   313   320   382   385
## Predicted 11.9  61.2  2.38 -0.622  4.779 -4.815  4.39  48.5  15.90  10.6
## cvpred    14.2  56.5 -6.25 -15.144 -0.601 -0.237  1.89  43.5  9.55 -15.0
## num_deaths 0.0   23.0  0.00 10.000 15.000  0.000  0.00  63.0  0.00   0.0
## CV residual -14.2 -33.5  6.25 25.144 15.601  0.237 -1.89 19.5 -9.55  15.0
##          434   550   628   630   647   680   761   808   881   936   94
4
## Predicted 22.08 42.8  47.4   8.72 11.51 187  -7.03  4.74  31.5 29.37 27.
2
## cvpred    -1.13  39.5  72.0  18.08  7.41 151   1.64 -5.08  53.0 20.29 20.
3
## num_deaths 25.00 21.0   0.0   0.00 52.00 440   0.00  0.00 10.0 17.00 33.
0
## CV residual 26.13 -18.5 -72.0 -18.08 44.59 289  -1.64  5.08 -43.0 -3.29 12.
7
##          962   963   965   992   1083  1136  1138  1171  1197  1221
1308

```

```

## Predicted    67.5 16.2 3.95 11.0 -7.03 -1.86 167.6 25.41 -3.203 -4.066
-13.1
## cvpred     73.3 12.1 -2.39 11.8 1.64 -8.21 136.4 18.85 0.604 0.843
-22.9
## num_deaths 0.0 0.0 0.00 0.0 0.00 0.00 42.0 12.00 0.000 0.000
0.0
## CV residual -73.3 -12.1 2.39 -11.8 -1.64 8.21 -94.4 -6.85 -0.604 -0.843
22.9
##          1315 1421 1461 1490 1506 1513 1544 1562 1626 1636 164
8
## Predicted   -7.779 15.9 147 14 178 11.65 -3.37 67.8 10.67 -14.0 48.
5
## cvpred      0.564 18.9 57 15 163 21.03 11.06 66.5 10.13 -15.2 43.
5
## num_deaths  0.000 67.0 268 10 42 15.00 0.00 180.0 16.00 0.0 17.
0
## CV residual -0.564 48.1 211 -5 -121 -6.03 -11.06 113.5 5.87 15.2 -26.
5
##          1775 1776 1814 1860 1883 1979 2012 2013 2019
## Predicted   -0.53 8.65 11.83 5.57 8.97 8.041 186 154.4 11.6
## cvpred      -2.36 10.61 4.16 -12.71 14.98 0.759 147 59.2 21.0
## num_deaths  0.00 0.00 0.00 0.00 0.00 22.000 668 464.0 0.0
## CV residual 2.36 -10.61 -4.16 12.71 -14.98 21.241 521 404.8 -21.0
##
## Sum of squares = 624630      Mean square = 12012      n = 52
##
## fold 5
## Observations in test set: 52
##          5 12 26 31 123 152 176 179 204 224
306
## Predicted   11.6 2.44 -2.55 -4.07 7.18 90.3 8.723 143.1 34.00 178 3
3.4
## cvpred      9.7 -17.60 -19.95 -5.11 8.07 104.3 -0.611 137.7 25.64 131 3
7.0
## num_deaths  0.0 0.00 0.00 0.00 0.00 48.0 15.000 160.0 34.00 509 1
0.0
## CV residual -9.7 17.60 19.95 5.11 -8.07 -56.3 15.611 22.3 8.36 378 -2
7.0
##          411 486 533 537 757 840 849 858 871 910
946
## Predicted   22.9 27.2 -4.81 0.109 11.2 15.9 25.4 55.5 16.22 211.8 4
.74
## cvpred      22.9 38.8 -7.37 -2.461 22.0 23.7 36.1 67.0 18.75 259.4 16
.50
## num_deaths  40.0 25.0 0.00 0.000 0.0 17.0 18.0 28.0 22.00 164.0 0
.00
## CV residual 17.1 -13.8 7.37 2.461 -22.0 -6.7 -18.1 -39.0 3.25 -95.4 -16
.50
##          949 954 985 1046 1090 1091 1207 1216 1287 1293
## Predicted   12.4 186.3 4.39 167.6 -1.86 18.43 -3.01 -2.55 -3.29 -0.0579

```

```

## cvpred      12.9 161.8  8.42 133.6  1.83 18.05 -5.38 -19.95 -3.68  0.9564
## num_deaths 14.0 143.0  0.00 172.0  0.00 20.00  0.00   0.00  0.00  0.0000
## CV residual 1.1 -18.8 -8.42  38.4 -1.83  1.95  5.38  19.95  3.68 -0.9564
##               1312 1362 1428 1433 1456 1772 1815 1866 1876 1879 192
9
## Predicted    8.48 -16.3 22.5 31.5 61.8  33.4  18.5  4.74  205  1.81 23.1
8
## cvpred      10.62 -18.7 17.2 10.0 65.3  37.0  31.1  16.50  261  4.36 23.7
7
## num_deaths   0.00   0.0 29.0 53.0 85.0  27.0   0.0   0.00  104  0.00 19.0
0
## CV residual -10.62  18.7 11.8 43.0 19.7 -10.0 -31.1 -16.50 -157 -4.36 -4.7
7
##               1957 2010 2062 2106 2117  2122  2140  2195 2289
## Predicted    7.18 16.84 48.47 205 18.8 7.11e-15 8.65 1.81 11.9
## cvpred      8.07 29.99 53.83 261 25.3 5.32e+01 7.73 4.36 10.6
## num_deaths   0.00 33.00 55.00 126 0.0 0.00e+00 0.00 0.00 0.0
## CV residual -8.07 3.01 1.17 -135 -25.3 -5.32e+01 -7.73 -4.36 -10.6
##
## Sum of squares = 213912      Mean square = 4114      n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5442

```

##Prune out *age45-54:hypertensive* ## Adjusted R-squared: 0.253 ## Average mean Square error 5427

```

covidcleansed$`age45-54SQ`<- covidcleansed$`age45-54`^2
model.interactionC <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothc
ondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` +
  alzheimer*`age45-54` + cardiacArrest*`age45-54` + cardiacarrhythmia*`age4
5-54` + cerebrovascula*`age45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` + heartfa
ilure*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-
54` + malignant*`age45-54` +
  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*
`age45-54` + renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + sepsis*`age45-54` + vascular*`age
45-54` + Northeast*`age45-54` +

```

```

  South*`age45-54` + West*`age45-54` , data = train)
summary(model.interactionC)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##      alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##      chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##      influenza + injuryPoisoning + ischemicheart + malignant +
##      obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##      conditionrespiratoryarrest + sepsis + vascular + Northeast +
##      South + West + allothcondsandcauses * `age45-54` + alzheimer *
##      `age45-54` + cardiacArrest * `age45-54` + cardiacarrhythmia *
##      `age45-54` + cerebrovascula * `age45-54` + chroniclr * `age45-54` +
##      covid19 * `age45-54` + diabetes * `age45-54` + heartfailure *
##      `age45-54` + influenza * `age45-54` + injuryPoisoning * `age45-54` +
##      ischemicheart * `age45-54` + malignant * `age45-54` + obesity *
##      `age45-54` + othercirculatorysystem * `age45-54` + otherrespsystem *
##      `age45-54` + renalfailure * `age45-54` + conditionrespiratoryarrest *
##      `age45-54` + sepsis * `age45-54` + vascular * `age45-54` +
##      Northeast * `age45-54` + South * `age45-54` + West * `age45-54` ,
##      data = train)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -195.9   -17.4   -4.5    4.2   980.9
##
## Coefficients:
## (Intercept)            37.88   13.35   2.84   0.0046
## `age45-54`              22.34   14.56   1.53   0.1254
## adultrespdistSyn      -27.99   13.97  -2.00   0.0453
## allothcondsandcauses     2.26   16.91   0.13   0.8939
## alzheimer             -43.06   15.92  -2.70   0.0070
## cardiacArrest           -28.72   19.74  -1.46   0.1459
## cardiacarrhythmia       -41.84   17.74  -2.36   0.0185
## cerebrovascula          -42.11   18.57  -2.27   0.0235
## chroniclr               -40.45   19.09  -2.12   0.0344
## covid19                  22.41   16.79   1.33   0.1824
## diabetes                 -16.97   21.41  -0.79   0.4282
## heartfailure             -41.11   19.75  -2.08   0.0377
## hypertensive              -30.70   13.73  -2.24   0.0255

```

|                                           |        |       |       |         |  |
|-------------------------------------------|--------|-------|-------|---------|--|
| *                                         |        |       |       |         |  |
| ## influenza                              | 1.88   | 17.54 | 0.11  | 0.9149  |  |
| ## injuryPoisoning                        | -39.00 | 18.33 | -2.13 | 0.0336  |  |
| *                                         |        |       |       |         |  |
| ## ischemicheart                          | -41.34 | 18.14 | -2.28 | 0.0229  |  |
| *                                         |        |       |       |         |  |
| ## malignant                              | -41.41 | 19.16 | -2.16 | 0.0309  |  |
| *                                         |        |       |       |         |  |
| ## obesity                                | -17.11 | 19.10 | -0.90 | 0.3706  |  |
| ## othercirculatorysystem                 | -37.78 | 19.09 | -1.98 | 0.0481  |  |
| *                                         |        |       |       |         |  |
| ## otherrespsystem                        | -40.12 | 19.76 | -2.03 | 0.0426  |  |
| *                                         |        |       |       |         |  |
| ## renalfailure                           | -29.02 | 22.53 | -1.29 | 0.1980  |  |
| ## conditionrespiratoryarrest             | -40.35 | 19.23 | -2.10 | 0.0361  |  |
| *                                         |        |       |       |         |  |
| ## sepsis                                 | -31.66 | 18.38 | -1.72 | 0.0852  |  |
| .                                         |        |       |       |         |  |
| ## vascular                               | -42.81 | 16.03 | -2.67 | 0.0077  |  |
| **                                        |        |       |       |         |  |
| ## Northeast                              | 4.37   | 9.78  | 0.45  | 0.6547  |  |
| ## South                                  | 9.45   | 8.61  | 1.10  | 0.2728  |  |
| ## West                                   | 4.88   | 8.62  | 0.57  | 0.5713  |  |
| ## `age45-54` :allothcondsandcauses       | 7.65   | 21.12 | 0.36  | 0.7171  |  |
| ## `age45-54` :alzheimer                  | -36.95 | 19.67 | -1.88 | 0.0606  |  |
| .                                         |        |       |       |         |  |
| ## `age45-54` :cardiacArrest              | -19.48 | 24.48 | -0.80 | 0.4264  |  |
| ## `age45-54` :cardiacarrhythmia          | -32.64 | 23.09 | -1.41 | 0.1578  |  |
| ## `age45-54` :cerebrovascula             | -32.75 | 25.07 | -1.31 | 0.1918  |  |
| ## `age45-54` :chroniclr                  | -26.87 | 23.60 | -1.14 | 0.2551  |  |
| ## `age45-54` :covid19                    | 93.89  | 20.47 | 4.59  | 5.1e-06 |  |
| ***                                       |        |       |       |         |  |
| ## `age45-54` :diabetes                   | -16.07 | 24.40 | -0.66 | 0.5104  |  |
| ## `age45-54` :heartfailure               | -35.88 | 29.59 | -1.21 | 0.2256  |  |
| ## `age45-54` :influenza                  | 47.93  | 21.88 | 2.19  | 0.0287  |  |
| *                                         |        |       |       |         |  |
| ## `age45-54` :injuryPoisoning            | -31.41 | 23.79 | -1.32 | 0.1870  |  |
| ## `age45-54` :ischemicheart              | -19.07 | 24.74 | -0.77 | 0.4409  |  |
| ## `age45-54` :malignant                  | -29.51 | 26.27 | -1.12 | 0.2614  |  |
| ## `age45-54` :obesity                    | -28.15 | 24.15 | -1.17 | 0.2442  |  |
| ## `age45-54` :othercirculatorysystem     | -26.16 | 25.43 | -1.03 | 0.3038  |  |
| ## `age45-54` :otherrespsystem            | -33.57 | 26.71 | -1.26 | 0.2092  |  |
| ## `age45-54` :renalfailure               | -32.28 | 25.85 | -1.25 | 0.2121  |  |
| ## `age45-54` :conditionrespiratoryarrest | -38.29 | 25.55 | -1.50 | 0.1343  |  |
| ## `age45-54` :sepsis                     | -10.40 | 22.70 | -0.46 | 0.6469  |  |
| ## `age45-54` :vascular                   | -35.17 | 21.05 | -1.67 | 0.0950  |  |
| .                                         |        |       |       |         |  |
| ## `age45-54` :Northeast                  | 25.04  | 13.46 | 1.86  | 0.0632  |  |
| .                                         |        |       |       |         |  |
| ## `age45-54` :South                      | 18.15  | 11.87 | 1.53  | 0.1266  |  |



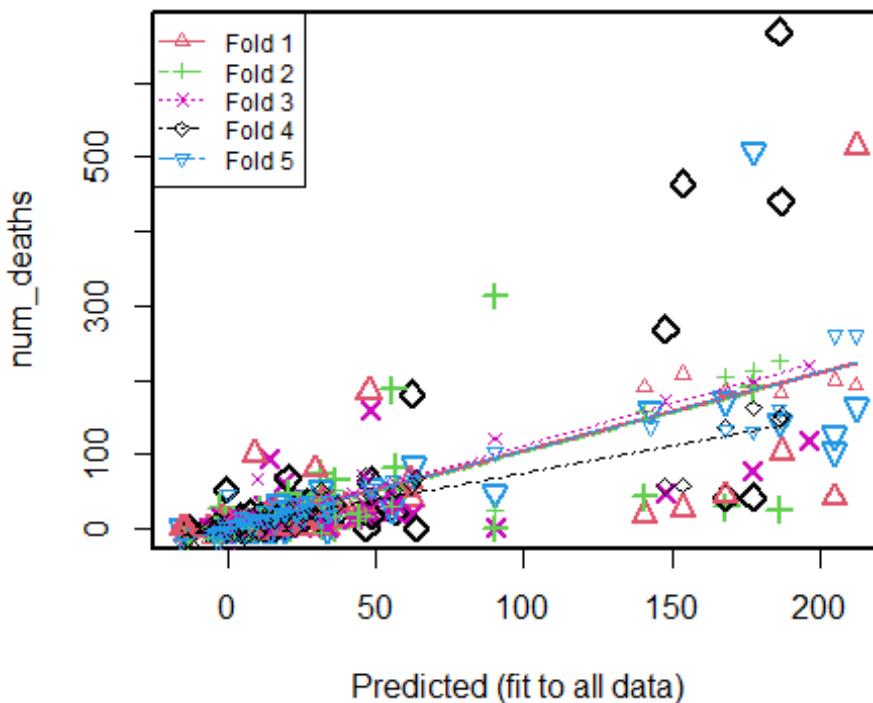
|                                           |   |        |        |       |         |
|-------------------------------------------|---|--------|--------|-------|---------|
| ## adultrespdistSyn                       | 1 | 308    | 308    | 0.06  | 0.80094 |
| ## allothcondsandcauses                   | 1 | 124916 | 124916 | 25.84 | 8.2e-07 |
| ***                                       |   |        |        |       |         |
| ## alzheimer                              | 1 | 17801  | 17801  | 3.68  | 0.05636 |
| .                                         |   |        |        |       |         |
| ## cardiacArrest                          | 1 | 9110   | 9110   | 1.88  | 0.17131 |
| ## cardiacarrhythmia                      | 1 | 1509   | 1509   | 0.31  | 0.57694 |
| ## cerebrovascula                         | 1 | 512    | 512    | 0.11  | 0.74517 |
| ## chroniclr                              | 1 | 1137   | 1137   | 0.24  | 0.62821 |
| ## covid19                                | 1 | 355946 | 355946 | 73.62 | 2.0e-15 |
| ***                                       |   |        |        |       |         |
| ## diabetes                               | 1 | 3533   | 3533   | 0.73  | 0.39364 |
| ## heartfailure                           | 1 | 499    | 499    | 0.10  | 0.74826 |
| ## hypertensive                           | 1 | 196    | 196    | 0.04  | 0.84051 |
| ## influenza                              | 1 | 7406   | 7406   | 1.53  | 0.21723 |
| ## injuryPoisoning                        | 1 | 992    | 992    | 0.21  | 0.65103 |
| ## ischemicheart                          | 1 | 500    | 500    | 0.10  | 0.74819 |
| ## malignant                              | 1 | 1759   | 1759   | 0.36  | 0.54708 |
| ## obesity                                | 1 | 5      | 5      | 0.00  | 0.97315 |
| ## othercirculatorysystem                 | 1 | 46     | 46     | 0.01  | 0.92241 |
| ## otherrespsystem                        | 1 | 203    | 203    | 0.04  | 0.83778 |
| ## renalfailure                           | 1 | 4668   | 4668   | 0.97  | 0.32694 |
| ## conditionrespiratoryarrest             | 1 | 1303   | 1303   | 0.27  | 0.60418 |
| ## sepsis                                 | 1 | 176    | 176    | 0.04  | 0.84905 |
| ## vascular                               | 1 | 12311  | 12311  | 2.55  | 0.11206 |
| ## Northeast                              | 1 | 113    | 113    | 0.02  | 0.87847 |
| ## South                                  | 1 | 3229   | 3229   | 0.67  | 0.41470 |
| ## West                                   | 1 | 2523   | 2523   | 0.52  | 0.47091 |
| ## `age45-54` :allothcondsandcauses       | 1 | 82454  | 82454  | 17.05 | 5.2e-05 |
| ***                                       |   |        |        |       |         |
| ## `age45-54` :alzheimer                  | 1 | 1892   | 1892   | 0.39  | 0.53233 |
| ## `age45-54` :cardiacArrest              | 1 | 2956   | 2956   | 0.61  | 0.43515 |
| ## `age45-54` :cardiacarrhythmia          | 1 | 618    | 618    | 0.13  | 0.72097 |
| ## `age45-54` :cerebrovascula             | 1 | 200    | 200    | 0.04  | 0.83919 |
| ## `age45-54` :chroniclr                  | 1 | 526    | 526    | 0.11  | 0.74185 |
| ## `age45-54` :covid19                    | 1 | 4981   | 4981   | 1.03  | 0.31128 |
| ## `age45-54` :diabetes                   | 1 | 600    | 600    | 0.12  | 0.72507 |
| ## `age45-54` :heartfailure               | 1 | 21     | 21     | 0.00  | 0.94796 |
| ## `age45-54` :influenza                  | 1 | 424    | 424    | 0.09  | 0.76737 |
| ## `age45-54` :injuryPoisoning            | 1 | 210    | 210    | 0.04  | 0.83500 |
| ## `age45-54` :ischemicheart              | 1 | 1065   | 1065   | 0.22  | 0.63933 |
| ## `age45-54` :malignant                  | 1 | 1067   | 1067   | 0.22  | 0.63904 |
| ## `age45-54` :obesity                    | 1 | 351    | 351    | 0.07  | 0.78800 |
| ## `age45-54` :othercirculatorysystem     | 1 | 2      | 2      | 0.00  | 0.98343 |
| ## `age45-54` :otherrespsystem            | 1 | 247    | 247    | 0.05  | 0.82132 |
| ## `age45-54` :renalfailure               | 1 | 2033   | 2033   | 0.42  | 0.51739 |
| ## `age45-54` :conditionrespiratoryarrest | 1 | 337    | 337    | 0.07  | 0.79211 |
| ## `age45-54` :sepsis                     | 1 | 54     | 54     | 0.01  | 0.91632 |
| ## `age45-54` :vascular                   | 1 | 3916   | 3916   | 0.81  | 0.36919 |
| ## `age45-54` :Northeast                  | 1 | 1784   | 1784   | 0.37  | 0.54427 |

```

## `age45-54` :South          1      24      24   0.00 0.94432
## `age45-54` :West          1     354     354   0.07 0.78709
## Residuals                 212 1025038    4835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
## :
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

## Small symbols show cross-validation predicted values



```

##
## fold 1
## Observations in test set: 52
##           21    27    38    79   186   280   359   396   413   418
499
## Predicted   -0.966  14.1   8.25   5.93   47.5  62.13   154   4.6  24.84 15.793
33.3
## cvpred      0.597  15.8   3.44   9.68   29.3  48.58   208  11.0  31.62 21.379
43.4
## num_deaths   0.000    0.0    0.00    0.00  185.0  39.00    26    0.0  30.00 21.000
0.0
## CV residual -0.597 -15.8 -3.44 -9.68 155.7 -9.58 -182 -11.0 -1.62 -0.379 -
43.4

```

```

##          509   544   578   609   668   675   699   726   728   810
818
## Predicted  1.29  30.2 -13.58 -3.21   8.60 -2.22 -4.29 187.1  30.2  2.14
205
## cvpred    3.87  39.9 -5.99 -2.66   5.34 -7.83 -8.62 182.6  39.9  7.39
200
## num_deaths 0.00  22.0  0.00  0.00 100.00 13.00  0.00 103.0  10.0  0.00
41
## CV residual -3.87 -17.9  5.99  2.66  94.66 20.83  8.62 -79.6 -29.9 -7.39
-159
##          917   972  1007  1015  1100  1127  1139  1174  1266  1413  14
64
## Predicted 24.84 10.80 13.31 -0.744  38.2   4.33  141  26.8  26.8  26.0 39.
47
## cvpred    31.62 -7.28  8.21  1.389  32.7  13.00 191  22.3  22.3  29.4 36.
45
## num_deaths 24.00  0.00 14.00 11.000  11.0  0.00  18  17.0  14.0  0.0 31.
00
## CV residual -7.62  7.28  5.79  9.611 -21.7 -13.00 -173 -5.3 -8.3 -29.4 -5.
45
##          1500  1517  1564 1586  1589  1600  1689  1892  1909  191
4
## Predicted -6.74 -0.724 22.642 61.8  0.602  212  31.9 -14.5  4.33 -16.0
4
## cvpred    -13.04  0.398 13.871 51.4 -3.362 192  42.8 -17.0 13.00 -9.5
7
## num_deaths 0.00  0.000 14.000 64.0  0.000  514  10.0  0.0  0.00  0.0
0
## CV residual 13.04 -0.398  0.129 12.6  3.362  322 -32.8  17.0 -13.00  9.5
7
##          1965 1994  2071 2138  2174 2236  2275 2304  2313
## Predicted 31.93 29.2  7.75 11.7 13.71 168  12.1 20.31  1.25
## cvpred    42.82 24.8 12.55  3.7  7.29 186  17.5  4.37  2.75
## num_deaths 39.00 79.0  0.00  0.0 15.00  44  0.0  0.00  0.00
## CV residual -3.82 54.2 -12.55 -3.7  7.71 -142 -17.5 -4.37 -2.75
##
## Sum of squares = 263273      Mean square = 5063      n = 52
##
## fold 2
## Observations in test set: 53
##          39   125   156   178   198   266   310   352   355   371
## Predicted 21.1   5.93  6.81 177.5  90.3 46.63  9.25  2.14 10.7  7.75
## cvpred    26.4  10.24 -6.26 212.8  24.0 17.15 17.02  2.26 17.6 11.48
## num_deaths 0.0   0.00 11.00 190.0 313.0 22.00  0.00  0.00  0.0  0.00
## CV residual -26.4 -10.24 17.26 -22.8 289.0  4.85 -17.02 -2.26 -17.6 -11.48
##          397   412   457   458   492   503   517   571   586   635
656
## Predicted  5.93 56.4  26.2 56.371 55.5  15.3 -3.21 -2.22 168 31.27 -2
.49
## cvpred    10.24 53.5 28.3 53.468 29.1  16.9 -1.83  2.62 203 -4.45 -15

```

```

.77
## num_deaths      0.00 82.0  15.0 53.000 42.0    0.0  0.00  0.00   29 74.00  28
.00
## CV residual -10.24 28.5 -13.3 -0.468 12.9 -16.9  1.83 -2.62 -174 78.45  43
.77
##          736   876   894   900   977   982  1003 1028 1034 1039   10
41
## Predicted -2.65 22.9 12.56 11.7  1.73  36.3 31.27 11.0 36.5 -6.95 -0.7
86
## cvpred     -6.82 32.7 20.88 18.5 -2.95  51.1 -4.45 14.9 16.6 -5.61 -6.0
87
## num_deaths  0.00 10.0 11.00  0.0  0.00  0.0 46.00 25.0 65.0  0.00 12.0
00
## CV residual 6.82 -22.7 -9.88 -18.5  2.95 -51.1 50.45 10.1 48.4  5.61 18.0
87
##          1084 1204 1222 1239 1288 1321 1511 1559 1567 1645 167
5
## Predicted -13.6 90.3 -4.27 24.8 -2.65 19.04 4.22 20.3 6.22 141 10.
7
## cvpred     -17.7 24.0 -8.22 27.6 -6.82 21.94 7.08 15.7 3.01 157 18.
5
## num_deaths  0.0  0.0  0.00 14.0  0.00 14.00 0.00 47.0 12.00 43  0.
0
## CV residual 17.7 -24.0  8.22 -13.6  6.82 -7.94 -7.08 31.3 8.99 -114 -18.
5
##          1690 1734 1773 1873 1983 2008 2035 2061 2097 2211
## Predicted  186  8.25  0.602 31.93 13.58 55.5 -3.21 44.2  5.93 -4.29
## cvpred     225  9.18 -3.686 37.79 13.63 29.1 -1.83 11.4 10.24 -7.92
## num_deaths  25  0.00  0.000 42.00 17.00 187.0 0.00 20.0  0.00  0.00
## CV residual -200 -9.18  3.686  4.21  3.37 157.9 1.83  8.6 -10.24  7.92
##
## Sum of squares = 215763      Mean square = 4071      n = 53
##
## fold 3
## Observations in test set: 53
##          16   190    209   211   274    307   325   428   454   4
87
## Predicted  90.3  6.65 -0.187 14.06 55.5  0.602 24.8 15.692 48.3  6.
52
## cvpred     120.3 16.51 -10.599 -3.83 52.3  8.520 25.4 17.857 38.7  9.
39
## num_deaths  0.0 14.00 13.000 93.00 23.0  0.000  0.0 17.000 158.0  0.
00
## CV residual -120.3 -2.51 23.599 96.83 -29.3 -8.520 -25.4 -0.857 119.3 -9.
39
##          491   527    597   677   774    793   868   904   907   909
967
## Predicted 12.10 14.06 -0.724 19.04 48.3  7.65 10.0 62.6 26.0 148  1
.83
## cvpred    15.68 -3.83 -4.818  6.63 38.7 14.05 66.4 70.7 25.1 172  8

```

```

.96
## num_deaths 12.00 0.00 0.000 61.00 18.0 0.00 10.0 21.0 40.0 48 0
.00
## CV residual -3.68 3.83 4.818 54.37 -20.7 -14.05 -56.4 -49.7 14.9 -124 -8
.96
## 994 1009 1014 1071 1213 1225 1327 1334 1375 1
391
## Predicted 9.25 11.95 -2.39 -3.00 -9.77e-14 1.25 2.19 -2.65 20.3 1
.73
## cvpred 12.83 6.01 -8.06 -2.02 -3.39e+01 -1.67 2.74 -6.19 29.8 13
.18
## num_deaths 0.00 32.00 15.00 0.00 0.00e+00 0.00 0.00 0.00 0.0 0
.00
## CV residual -12.83 25.99 23.06 2.02 3.39e+01 1.67 -2.74 6.19 -29.8 -13
.18
## 1438 1472 1523 1540 1554 1760 1805 1818 1836 1838
## Predicted 22.8 6.65 2.727 45.8 196.4 22.8 1.73 33.9 56.4 55.02
## cvpred 31.4 16.51 0.836 70.9 218.4 31.4 13.18 43.1 61.4 56.17
## num_deaths 13.0 0.00 0.000 15.0 119.0 11.0 0.00 0.0 33.0 49.00
## CV residual -18.4 -16.51 -0.836 -55.9 -99.4 -20.4 -13.18 -43.1 -28.4 -7.17
## 1936 1960 2041 2050 2052 2125 2152 2190 2225
2248
## Predicted 42.3 2.140 -9.77e-14 -4.27 -6.74 3.96 47.5 178 4.33e+00
15.8
## cvpred 47.2 -0.669 -3.39e+01 4.42 2.97 16.47 65.1 197 2.84e-14
11.0
## num_deaths 13.0 0.000 0.00e+00 0.00 0.00 0.00 25.0 78 1.30e+01
0.0
## CV residual -34.2 0.669 3.39e+01 -4.42 -2.97 -16.47 -40.1 -119 1.30e+01
-11.0
## 2252 2255
## Predicted 3.69 8.59
## cvpred 2.78 19.36
## num_deaths 0.00 0.00
## CV residual -2.78 -19.36
##
## Sum of squares = 104721 Mean square = 1976 n = 53
##
## fold 4
## Observations in test set: 52
## 7 94 126 159 183 217 313 320 382 385
434
## Predicted 14.0 56.4 2.14 -0.425 4.220 -4.918 4.74 48.3 15.7 10.43 2
2.586
## cvpred 16.3 51.5 -6.27 -5.788 -0.727 0.269 2.80 43.5 9.5 -6.43 -
0.478
## num_deaths 0.0 23.0 0.00 10.000 15.000 0.000 0.00 63.0 0.0 0.00 2
5.000
## CV residual -16.3 -28.5 6.27 15.788 15.727 -0.269 -2.80 19.5 -9.5 6.43 2
5.478

```

```

##          550   628   630   647   680   761   808   881   936   944   9
62
## Predicted  36.8  46.6   8.25 -0.744  187  -6.36   4.60  31.5 29.19 26.8  63
.5
## cvpred    32.5  71.3  17.78 -3.996  152   1.86  -5.06  53.0 20.31 20.2  68
.7
## num_deaths 21.0   0.0   0.00 52.000  440   0.00   0.00 10.0 17.00 33.0   0
.0
## CV residual -11.5 -71.3 -17.78 55.996  288  -1.86   5.06 -43.0 -3.31 12.8 -68
.7
##          963   965   992  1083  1136  1138  1171  1197  1221  1308
1315
## Predicted  18.9  4.21  11.7  -6.36 -1.06 168.2 24.91 -3.096 -4.328 -12.7
-6.95
## cvpred     14.9 -1.75  12.2   1.86 -7.64 136.7 18.04  0.907  0.889 -22.3
1.24
## num_deaths  0.0   0.00   0.0   0.00  0.00  42.0 12.00  0.000  0.000   0.0
0.00
## CV residual -14.9  1.75 -12.2  -1.86  7.64 -94.7 -6.04 -0.907 -0.889  22.3
-1.24
##          1421  1461  1490  1506  1513  1544  1562  1626  1636  1648   1
775
## Predicted  20.3 147.7 13.7   178  15.3  -4.27  62.1 11.01 -13.6  48.3  0.0
116
## cvpred     24.0  57.6 15.0   162  25.4  10.42  59.6 10.37 -15.0  43.5 -1.1
591
## num_deaths 67.0 268.0 10.0   42  15.0   0.00 180.0 16.00   0.0  17.0  0.0
000
## CV residual 43.0 210.4 -5.0  -120 -10.4 -10.42 120.4   5.63  15.0 -26.5  1.1
591
##          1776  1814  1860  1883  1979  2012  2013  2019
## Predicted  9.25 12.56  5.45  12.0  7.754 186 153.6 15.3
## cvpred    10.95  4.83 -12.39 17.3  0.261 147  58.4 25.4
## num_deaths  0.00  0.00   0.00   0.0 22.000 668 464.0  0.0
## CV residual -10.95 -4.83 12.39 -17.3 21.739 521 405.6 -25.4
##
## Sum of squares = 625603      Mean square = 12031      n = 52
##
## fold 5
## Observations in test set: 52
##          5    12    26    31   123   152    176    179    204   224    3
06
## Predicted 15.3  33.4  -3.43 -4.33  6.52  90.3  8.247 142.8 34.0 178  33
.9
## cvpred    13.0  27.7 -19.73 -5.31  7.48 104.3 -0.615 137.5 19.2 131  38
.0
## num_deaths  0.0   0.0   0.00  0.00  0.00  48.0 15.000 160.0 34.0 509  10
.0
## CV residual -13.0 -27.7 19.73  5.31 -7.48 -56.3 15.615 22.5 14.8 378 -28
.0

```

```

##          411   486   533   537   757   840   849   858   871   910
946
## Predicted  26.2  26.8 -4.92 -0.187  12.0 15.69  24.9  55.5 18.92 212.4
4.6
## cvpred    25.8  38.2 -7.39 -2.623  22.9 22.77  35.8  66.2 20.87 259.8  1
5.4
## num_deaths 40.0  25.0  0.00  0.000   0.0 17.00  18.0  28.0 22.00 164.0
0.0
## CV residual 14.2 -13.2  7.39  2.623 -22.9 -5.77 -17.8 -38.2  1.13 -95.8 -1
5.4
##          949   954   985  1046  1090  1091  1207  1216  1287  1293
1312
## Predicted 12.10 186.4  4.74  168 -1.06 19.04 -3.21 -3.43 -2.76  0.695
8.6
## cvpred    12.49 160.8  8.30  134  1.85 18.83 -5.34 -19.73 -2.84  2.003
11.4
## num_deaths 14.00 143.0  0.00  172  0.00 20.00  0.00  0.00  0.00  0.000
0.0
## CV residual 1.51 -17.8 -8.30   38 -1.85  1.17  5.34  19.73  2.84 -2.003 -
11.4
##          1362  1428  1433  1456  1772  1815  1866  1876  1879  1929  1957
2010
## Predicted -16.0 22.9 31.5 62.6  33.9 19.0   4.6  205  2.19 24.84  6.52
17.1
## cvpred    -18.2 18.0 10.0 65.9  38.0 31.2  15.4  260  4.93 25.47  7.48
28.7
## num_deaths  0.0 29.0 53.0 85.0  27.0  0.0   0.0  104  0.00 19.00  0.00
33.0
## CV residual 18.2 11.0 43.0 19.1 -11.0 -31.2 -15.4 -156 -4.93 -6.47 -7.48
4.3
##          2062  2106  2117   2122  2140  2195  2289
## Predicted 48.35  205   6.22 -1.07e-14  9.25  2.19  14.0
## cvpred    53.16  260  13.94  4.68e+01  8.38  4.93  12.7
## num_deaths 55.00  126   0.00  0.00e+00  0.00  0.00   0.0
## CV residual 1.84 -134 -13.94 -4.68e+01 -8.38 -4.93 -12.7
##
## Sum of squares = 212539      Mean square = 4087      n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5427

```

Pruned *age45-54:sepsis*

Adjusted R-squared: 0.254

MS 5428

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionD <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothon
andsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +

```

```

chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West +
##
allothcondsandcauses*`age45-54` +
alzheimer*`age45-54` + cardiacArrest*`age45-54` + cardiacarrhythmia*`age45-54` +
cerebrovascula*`age45-54` +
chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` + heartfailure*`age45-54` +
influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-54` +
malignant*`age45-54` +
obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*`age45-54` +
renalfailure*`age45-54` +
conditionrespiratoryarrest*`age45-54` + vascular*`age45-54` + Northeast*`age45-54` +
South*`age45-54` + West*`age45-54` , data = train)
summary(model.interactionD)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
##     alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##     chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##     influenza + injuryPoisoning + ischemicheart + malignant +
##     obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##     conditionrespiratoryarrest + sepsis + vascular + Northeast +
##     South + West + allothcondsandcauses * `age45-54` + alzheimer *
##     `age45-54` + cardiacArrest * `age45-54` + cardiacarrhythmia *
##     `age45-54` + cerebrovascula * `age45-54` + chroniclr * `age45-54` +
##     covid19 * `age45-54` + diabetes * `age45-54` + heartfailure *
##     `age45-54` + influenza * `age45-54` + injuryPoisoning * `age45-54` +
##     ischemicheart * `age45-54` + malignant * `age45-54` + obesity *
##     `age45-54` + othercirculatorysystem * `age45-54` + otherrespsystem *
##     `age45-54` + renalfailure * `age45-54` + conditionrespiratoryarrest *
##     `age45-54` + vascular * `age45-54` + Northeast * `age45-54` +
##     South * `age45-54` + West * `age45-54` , data = train)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -195.9   -17.7   -4.5    4.2   980.9
##
## Coefficients:
## (Intercept)             Estimate Std. Error t value Pr(>|t|)
## (Intercept)            39.542     12.845    3.08   0.0021
## `age45-54`              19.617     13.296    1.48   0.1404
## adultrespdistSyn       -28.102     13.961   -2.01   0.0444

```

|                                     |         |        |       |         |  |
|-------------------------------------|---------|--------|-------|---------|--|
| *                                   |         |        |       |         |  |
| ## allothcondsandcauses             | 0.581   | 16.504 | 0.04  | 0.9719  |  |
| ## alzheimer                        | -44.728 | 15.496 | -2.89 | 0.0040  |  |
| **                                  |         |        |       |         |  |
| ## cardiacArrest                    | -30.377 | 19.396 | -1.57 | 0.1176  |  |
| ## cardiacarrhythmia                | -43.495 | 17.361 | -2.51 | 0.0124  |  |
| *                                   |         |        |       |         |  |
| ## cerebrovascula                   | -43.773 | 18.207 | -2.40 | 0.0164  |  |
| *                                   |         |        |       |         |  |
| ## chroniclr                        | -42.125 | 18.735 | -2.25 | 0.0248  |  |
| *                                   |         |        |       |         |  |
| ## covid19                          | 20.740  | 16.388 | 1.27  | 0.2059  |  |
| ## diabetes                         | -18.682 | 21.079 | -0.89 | 0.3757  |  |
| ## heartfailure                     | -42.764 | 19.413 | -2.20 | 0.0278  |  |
| *                                   |         |        |       |         |  |
| ## hypertensive                     | -30.674 | 13.722 | -2.24 | 0.0256  |  |
| *                                   |         |        |       |         |  |
| ## influenza                        | 0.200   | 17.149 | 0.01  | 0.9907  |  |
| ## injuryPoisoning                  | -40.667 | 17.952 | -2.27 | 0.0237  |  |
| *                                   |         |        |       |         |  |
| ## ischemicheart                    | -43.001 | 17.775 | -2.42 | 0.0157  |  |
| *                                   |         |        |       |         |  |
| ## malignant                        | -43.069 | 18.806 | -2.29 | 0.0222  |  |
| *                                   |         |        |       |         |  |
| ## obesity                          | -18.792 | 18.737 | -1.00 | 0.3161  |  |
| ## othercirculatorysystem           | -39.448 | 18.731 | -2.11 | 0.0354  |  |
| *                                   |         |        |       |         |  |
| ## otherrespystem                   | -41.773 | 19.425 | -2.15 | 0.0317  |  |
| *                                   |         |        |       |         |  |
| ## renalfailure                     | -30.699 | 22.220 | -1.38 | 0.1674  |  |
| ## conditionrespiratoryarrest       | -42.028 | 18.873 | -2.23 | 0.0262  |  |
| *                                   |         |        |       |         |  |
| ## sepsis                           | -37.393 | 13.453 | -2.78 | 0.0055  |  |
| **                                  |         |        |       |         |  |
| ## vascular                         | -44.490 | 15.599 | -2.85 | 0.0044  |  |
| **                                  |         |        |       |         |  |
| ## Northeast                        | 4.526   | 9.767  | 0.46  | 0.6431  |  |
| ## South                            | 9.479   | 8.611  | 1.10  | 0.2712  |  |
| ## West                             | 4.796   | 8.613  | 0.56  | 0.5778  |  |
| ## `age45-54` :allothcondsandcauses | 10.411  | 20.238 | 0.51  | 0.6070  |  |
| ## `age45-54` :alzheimer            | -34.211 | 18.735 | -1.83 | 0.0681  |  |
| .                                   |         |        |       |         |  |
| ## `age45-54` :cardiacArrest        | -16.736 | 23.725 | -0.71 | 0.4807  |  |
| ## `age45-54` :cardiacarrhythmia    | -29.914 | 22.304 | -1.34 | 0.1801  |  |
| ## `age45-54` :cerebrovascula       | -30.009 | 24.339 | -1.23 | 0.2179  |  |
| ## `age45-54` :chroniclr            | -24.127 | 22.820 | -1.06 | 0.2906  |  |
| ## `age45-54` :covid19              | 96.630  | 19.573 | 4.94  | 9.2e-07 |  |
| ***                                 |         |        |       |         |  |
| ## `age45-54` :diabetes             | -13.276 | 23.621 | -0.56 | 0.5742  |  |
| ## `age45-54` :heartfailure         | -33.153 | 28.972 | -1.14 | 0.2527  |  |

```

## `age45-54` : influenza          50.685   21.034    2.41   0.0161
*
## `age45-54` : injuryPoisoning   -28.660   23.009   -1.25   0.2132
## `age45-54` : ischemicheart     -16.333   23.995   -0.68   0.4962
## `age45-54` : malignant         -26.772   25.567   -1.05   0.2953
## `age45-54` : obesity           -25.391   23.385   -1.09   0.2778
## `age45-54` : othercirculatorysystem -23.417   24.704   -0.95   0.3434
## `age45-54` : otherrespsystem    -30.846   26.034   -1.18   0.2364
## `age45-54` : renalfailure       -29.527   25.133   -1.17   0.2403
## `age45-54` : conditionrespiratoryarrest -35.531   24.823   -1.43   0.1526
## `age45-54` : vascular          -32.422   20.168   -1.61   0.1082
## `age45-54` : Northeast          24.834   13.451    1.85   0.0651
.
## `age45-54` : South              18.085   11.867    1.52   0.1278
## `age45-54` : West               17.998   12.469    1.44   0.1492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.2 on 1053 degrees of freedom
## Multiple R-squared:  0.286, Adjusted R-squared:  0.254
## F-statistic:  8.8 on 48 and 1053 DF,  p-value: <2e-16

outD <- cv.lm(data = test
                , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses * `age45-54` +
  alzheimer * `age45-54` + cardiacArrest * `age45-54` + cardiacarrhythmia * `age45-54` +
  cerebrovascula * `age45-54` +
  chroniclr * `age45-54` + covid19 * `age45-54` + diabetes * `age45-54` + heartfailure * `age45-54` +
  influenza * `age45-54` + injuryPoisoning * `age45-54` + ischemicheart * `age45-54` +
  malignant * `age45-54` +
  obesity * `age45-54` + othercirculatorysystem * `age45-54` + otherrespsystem * `age45-54` +
  renalfailure * `age45-54` +
  conditionrespiratoryarrest * `age45-54` + vascular * `age45-54` + Northeast * `age45-54` +
  South * `age45-54` + West * `age45-54`)
                , plotit = "Observed", m=5)

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a
## rank-deficient fit may be misleading

```

```

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Analysis of Variance Table
##
## Response: num_deaths
##                                     Df Sum Sq Mean Sq F value Pr(>F)
## `age45-54`                         1 55638 55638 11.56 0.00081
## ***
## adultrespdistSyn                  1     308     308   0.06 0.80051
## allothcondsandcauses              1 124916 124916 25.95 7.7e-07
## ***
## alzheimer                          1 17801 17801  3.70 0.05582
## .
## cardiacArrest                      1  9110  9110  1.89 0.17037
## cardiacarrhythmia                 1  1509  1509  0.31 0.57611
## cerebrovascula                     1    512    512  0.11 0.74463
## chroniclr                           1   1137   1137  0.24 0.62746
## covid19                            1 355946 355946 73.94 1.8e-15
## ***
## diabetes                           1   3533   3533  0.73 0.39261
## heartfailure                       1     499    499  0.10 0.74772
## hypertensive                        1     196    196  0.04 0.84017
## influenza                           1   7406   7406  1.54 0.21622
## injuryPoisoning                    1    992    992  0.21 0.65031
## ischemicheart                      1    500    500  0.10 0.74766
## malignant                           1   1759   1759  0.37 0.54620
## obesity                             1      5      5  0.00 0.97310
## othercirculatorysystem              1     46     46  0.01 0.92224
## otherrespsystem                     1   203   203  0.04 0.83743
## renalfailure                        1   4668   4668  0.97 0.32588
## conditionrespiratoryarrest         1   1303   1303  0.27 0.60339
## sepsis                             1   176   176  0.04 0.84873
## vascular                            1 12311 12311  2.56 0.11127
## Northeast                           1    113    113  0.02 0.87821
## South                               1   3229   3229  0.67 0.41368
## West                                1   2523   2523  0.52 0.46993
## `age45-54`:allothcondsandcauses  1   82454  82454 17.13 5.0e-05
## ***
## `age45-54`:alzheimer               1   1892   1892  0.39 0.53143
## `age45-54`:cardiacArrest            1   2956   2956  0.61 0.43415
## `age45-54`:cardiacarrhythmia       1   618    618  0.13 0.72039
## `age45-54`:cerebrovascula           1   200    200  0.04 0.83884
## `age45-54`:chroniclr                1   526    526  0.11 0.74130
## `age45-54`:covid19                 1   4981   4981  1.03 0.31022

```

```

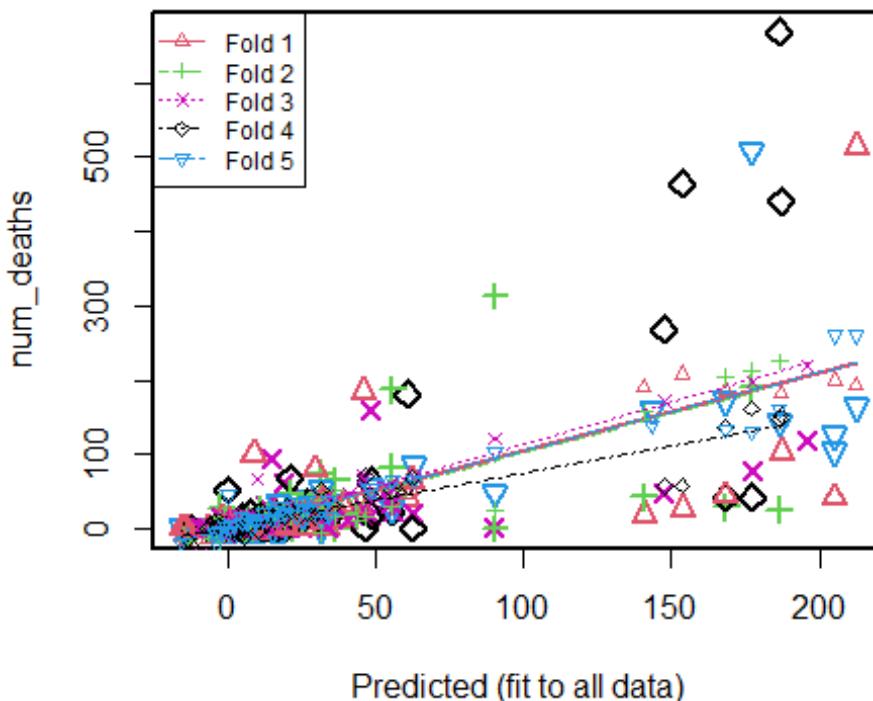
## `age45-54` :diabetes 1 600 600 0.12 0.72449
## `age45-54` :heartfailure 1 21 21 0.00 0.94784
## `age45-54` :influenza 1 424 424 0.09 0.76687
## `age45-54` :injuryPoisoning 1 210 210 0.04 0.83465
## `age45-54` :ischemicheart 1 1065 1065 0.22 0.63859
## `age45-54` :malignant 1 1067 1067 0.22 0.63830
## `age45-54` :obesity 1 351 351 0.07 0.78755
## `age45-54` :othercirculatorysystem 1 2 2 0.00 0.98339
## `age45-54` :otherrespsystem 1 247 247 0.05 0.82094
## `age45-54` :renalfailure 1 2033 2033 0.42 0.51647
## `age45-54` :conditionrespiratoryarrest 1 337 337 0.07 0.79167
## `age45-54` :vascular 1 3562 3562 0.74 0.39067
## `age45-54` :Northeast 1 1914 1914 0.40 0.52897
## `age45-54` :South 1 39 39 0.01 0.92796
## `age45-54` :West 1 268 268 0.06 0.81384
## Residuals 213 1025384 4814

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
:
## 
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

### Small symbols show cross-validation predicted val



```

## 
## fold 1
## Observations in test set: 52
##          21    27    38    79   186   280   359   396   413   418
499
## Predicted -0.558  14.4   8.06   5.81   45.9  60.75  154   4.66  25.26 15.731
33.7
## cvpred     0.989  16.0   3.34   9.60   28.2  47.66  208  11.01 31.88 21.325
43.8
## num_deaths  0.000   0.0   0.00   0.00 185.0  39.00   26   0.00 30.00 21.000
0.0
## CV residual -0.989 -16.0  -3.34  -9.60 156.8  -8.66  -182 -11.01 -1.88 -0.325
-43.8
##          509    544    578    609    668    675    699    726    728    810
818
## Predicted  2.42   30.2  -13.49 -2.89   8.59  -2.34  -4.29 187.1  30.2   2.16
205
## cvpred     4.80   39.9  -5.93  -2.42   5.31  -7.91  -8.61 182.6  39.9   7.36
200
## num_deaths  0.00   22.0   0.00   0.00 100.00  13.00   0.00 103.0  10.0   0.00
41
## CV residual -4.80  -17.9   5.93   2.42  94.69  20.91   8.61 -79.6  -29.9  -7.36
-159
##          917    972   1007   1015   1100   1127   1139   1174   1266   1413
1464
## Predicted 25.26  10.88  14.10 -0.185  37.1   4.33   141  26.73  26.73  25.9  3
9.09
## cvpred     31.88  -7.18   8.66   1.978  31.9   13.00  191  22.25  22.25  29.2  3
6.25
## num_deaths 24.00   0.00  14.00 11.000  11.0   0.00   18 17.00  14.00   0.0  3
1.00
## CV residual -7.88   7.18   5.34   9.022 -20.9  -13.00  -173 -5.25  -8.25  -29.2  -
5.25
##          1500   1517   1564  1586   1589  1600   1689  1892  1909   19
14
## Predicted -7.12  -0.418 22.8657 62.0   0.432  213   31.9  -14.5   4.33  -15.
99
## cvpred     -13.25  0.659 14.0252 51.6  -3.511  192   42.8  -16.9  13.00  -9.
57
## num_deaths  0.00   0.000 14.0000 64.0   0.000  514   10.0   0.0   0.00   0.
00
## CV residual 13.25 -0.659 -0.0252 12.4   3.511  322  -32.8  16.9  -13.00  9.
57
##          1965  1994   2071   2138   2174  2236   2275  2304  2313
## Predicted 31.87  29.2   7.65  11.90 19.84  168   11.9 19.94  1.63
## cvpred     42.77  24.8   12.48  3.83 11.97  186   17.4  4.15  3.01
## num_deaths 39.00  79.0   0.00   0.00 15.00   44   0.0  0.00  0.00
## CV residual -3.77  54.2  -12.48 -3.83   3.03  -142 -17.4  -4.15 -3.01
##
## Sum of squares = 263459      Mean square = 5067      n = 52

```

```

## 
## fold 2
## Observations in test set: 53
##          39     125     156     178     198     266     310     352     355     371
## Predicted 21.6    5.81    6.43   177.2    90.3   46.30     9.4    2.16   10.5    7.65
## cvpred   26.9   10.07   -6.59   212.6   24.0   16.88   17.2    2.25   17.5   11.38
## num_deaths 0.0    0.00   11.00  190.0  313.0   22.00     0.0    0.00    0.0    0.00
## CV residual -26.9  -10.07  17.59  -22.6   289.0    5.12  -17.2  -2.25  -17.5  -11.38
##          397    412    457    458    492    503    517    571    586    635
656
## Predicted      5.81  55.2   27.0  55.213  55.6   16.7  -2.89  -2.34   168  31.13   -2
.44
## cvpred       10.07  52.6   28.9  52.609  29.1   17.9  -1.55   2.37   203  -4.94  -15
.76
## num_deaths    0.00  82.0   15.0  53.000  42.0    0.0    0.00    0.00    29  74.00   28
.00
## CV residual  -10.07  29.4  -13.9   0.391  12.9  -17.9   1.55  -2.37  -174  78.94   43
.76
##          736    876    894    900    977    982   1003   1028   1034   1039    1
041
## Predicted    -2.52  23.0   12.8   11.9    1.39   36.5  31.13  11.08  36.6  -7.06  -0.
967
## cvpred      -6.66  32.8   21.1   18.6   -3.15   51.3  -4.94  15.01  16.8  -5.82  -6.
309
## num_deaths   0.00  10.0   11.0    0.0    0.00    0.0  46.00  25.00  65.0   0.00  12.
000
## CV residual   6.66  -22.8  -10.1  -18.6    3.15  -51.3  50.94   9.99  48.2   5.82  18.
309
##          1084   1204   1222   1239   1288   1321   1511   1559   1567   1645   167
5
## Predicted   -13.5  90.3  -4.62   25.3  -2.52  19.01   4.56  21.0   6.69   141   10.
5
## cvpred      -17.7  24.0  -8.48   28.0  -6.66  21.93   7.45  16.3   3.45   157   18.
3
## num_deaths   0.0   0.0   0.00   14.0   0.00  14.00   0.00  47.0  12.00    43   0.
0
## CV residual  17.7  -24.0   8.48  -14.0   6.66  -7.93  -7.45  30.7   8.55  -114  -18.
3
##          1690   1734   1773   1873   1983   2008   2035   2061   2097   2211
## Predicted   187   8.06   0.432  31.87  13.38   55.6  -2.89  44.00   5.81  -4.29
## cvpred     225   9.02  -3.803  37.82  13.42   29.1  -1.55  10.95  10.07  -7.94
## num_deaths  25   0.00   0.000  42.00  17.00  187.0   0.00  20.00    0.00   0.00
## CV residual -200  -9.02  3.803   4.18   3.58  157.9   1.55   9.05  -10.07   7.94
##
## Sum of squares = 216067      Mean square = 4077      n = 53
##
## fold 3
## Observations in test set: 53
##          16     190     209     211     274     307     325     428     454     4
87

```

```

## Predicted      90.3  6.35   0.265 14.40  55.6  0.432  25.3 15.705 48.4  6.
42
## cvpred       120.3 16.51 -10.599 -3.83  52.3  8.520  25.4 17.857 38.7  9.
39
## num_deaths    0.0 14.00 13.000 93.00  23.0  0.000  0.0 17.000 158.0  0.
00
## CV residual -120.3 -2.51  23.599 96.83 -29.3 -8.520 -25.4 -0.857 119.3 -9.
39
##               491   527   597   677   774   793   868   904   907   909
967
## Predicted     11.90 14.40 -0.418 19.01  48.4   7.38  10.0  62.8 25.9  148  1
.66
## cvpred        15.68 -3.83 -4.818  6.63  38.7  14.05  66.4  70.7 25.1  172  8
.96
## num_deaths    12.00  0.00  0.000 61.00  18.0   0.00  10.0  21.0 40.0  48  0
.00
## CV residual -3.68  3.83  4.818 54.37 -20.7 -14.05 -56.4 -49.7 14.9 -124 -8
.96
##               994  1009  1014  1071  1213  1225  1327  1334  1375  1391
1438
## Predicted     9.4 12.39 -2.41 -3.16 -9.4   1.63  1.96 -2.52 21.0  1.39
22.9
## cvpred        12.8  6.01 -8.06 -2.02 -33.9 -1.67  2.74 -6.19 29.8 13.18
31.4
## num_deaths    0.0 32.00 15.00  0.00   0.0  0.00  0.00  0.00  0.0  0.00
13.0
## CV residual -12.8 25.99 23.06  2.02  33.9  1.67 -2.74  6.19 -29.8 -13.18 -
18.4
##               1472  1523  1540  1554  1760  1805  1818  1836  1838  1936
## Predicted     6.35 3.112 45.4 196.0  22.9   1.39  34.0  55.2 53.51 40.9
## cvpred        16.51 0.836 70.9 218.4  31.4  13.18  43.1  61.4 56.17 47.2
## num_deaths    0.00 0.000 15.0 119.0  11.0   0.00  0.0  33.0 49.00 13.0
## CV residual -16.51 -0.836 -55.9 -99.4 -20.4 -13.18 -43.1 -28.4 -7.17 -34.2
##               1960  2041  2050  2052  2125  2152  2190   2225  2248  225
2
## Predicted     2.155 -9.4 -4.62 -7.12   3.72  45.9  177 4.33e+00 15.7  3.6
3
## cvpred        -0.669 -33.9  4.42  2.97  16.47  65.1  197 2.84e-14 11.0  2.7
8
## num_deaths    0.000  0.0  0.00  0.00   0.00  25.0   78 1.30e+01  0.0  0.0
0
## CV residual  0.669  33.9 -4.42 -2.97 -16.47 -40.1 -119 1.30e+01 -11.0 -2.7
8
##               2255
## Predicted     8.58
## cvpred        19.36
## num_deaths    0.00
## CV residual -19.36
##
## Sum of squares = 104721      Mean square = 1976      n = 53

```

```

## 
## fold 4
## Observations in test set: 52
##          7   94  126   159   183   217   313   320   382   385
434
## Predicted    15.0  55.2   2.16 -0.133   4.560 -4.459   4.54  48.4 15.70 10.13
29.12
## cvpred     17.4  50.5  -6.01 -2.712  -0.252   0.862   2.45  43.7  9.73 -4.09
15.68
## num_deaths   0.0  23.0   0.00 10.000  15.000   0.000   0.00  63.0   0.00   0.00
25.00
## CV residual -17.4 -27.5   6.01 12.712  15.252  -0.862  -2.45 19.3  -9.73  4.09
9.32
##          550   628   630   647   680   761   808   881   936   944   9
62
## Predicted   35.37  46.3   8.06 -0.185  187  -6.44   4.66  31.5 29.22 26.7   62
.5
## cvpred     30.56  70.7  17.49 -3.192  152   1.88  -4.78  53.0 20.62 20.3   67
.5
## num_deaths  21.00   0.0   0.00 52.000  440   0.00   0.00 10.0 17.00 33.0   0
.0
## CV residual -9.56 -70.7 -17.49 55.192  288  -1.88   4.78 -43.0 -3.62 12.7  -67
.5
##          963   965   992  1083  1136  1138  1171  1197  1221  1308
1315
## Predicted   19.3   3.86  11.9  -6.44 -0.799 168.4 24.67 -2.61 -3.84 -12.6  -7.06
## cvpred     15.2  -2.35  12.3   1.88 -7.318 136.9 17.67  1.66  1.55 -21.9
1.20
## num_deaths   0.0   0.00   0.0   0.00  0.000  42.0 12.00   0.00   0.00   0.0
0.00
## CV residual -15.2   2.35 -12.3  -1.88  7.318 -94.9 -5.67 -1.66 -1.55 21.9  -1.20
##          1421  1461  1490  1506  1513  1544  1562  1626  1636  1648   17
75
## Predicted   21.0  147.6  19.8   177   16.7  -4.62  60.8 11.08 -13.5  48.4  -0.1
86
## cvpred     24.9  57.4  30.3   162   27.1   9.85  57.8 10.45 -15.0  43.7  -1.3
72
## num_deaths  67.0 268.0  10.0   42   15.0   0.00 180.0 16.00   0.0  17.0   0.0
00
## CV residual 42.1 210.6 -20.3  -120  -12.1  -9.85 122.2  5.55  15.0  -26.7  1.3
72
##          1776  1814  1860  1883  1979  2012  2013  2019
## Predicted   9.4 12.78  5.54  12.4  7.652  187 153.6  16.7
## cvpred     11.1  5.29 -11.78  17.7  0.286  147  58.2  27.1
## num_deaths   0.0   0.00   0.00   0.0 22.000  668 464.0   0.0
## CV residual -11.1 -5.29  11.78 -17.7 21.714  521 405.8 -27.1
##
## Sum of squares = 625131      Mean square = 12022      n = 52

```

```

## 
## fold 5
## Observations in test set: 52
##          5   12   26   31   123   152   176   179   204   224   3
## 06
## Predicted    16.7  31.7 -3.74 -3.84  6.42  90.3  8.064 143.3 34.0 177  34
## .0
## cvpred      14.5  25.7 -20.27 -4.69  7.43 104.3 -0.737 138.3 16.8 131  38
## .1
## num_deaths   0.0   0.0   0.00  0.00  0.00  48.0 15.000 160.0 34.0 509  10
## .0
## CV residual -14.5 -25.7  20.27  4.69 -7.43 -56.3 15.737 21.7 17.2 378 -28
## .1
##          411   486   533   537   757   840   849   858   871   910
## 946
## Predicted    27.0  26.7 -4.46  0.265 11.8 15.70  24.7  55.6 19.269 213  4
## .66
## cvpred      26.6  38.2 -6.78 -1.913 22.4 22.82  35.3  66.3 21.324 260  15
## .48
## num_deaths   40.0  25.0  0.00  0.000  0.0 17.00  18.0  28.0 22.000 164  0
## .00
## CV residual 13.4 -13.2  6.78  1.913 -22.4 -5.82 -17.3 -38.3  0.676 -96 -15
## .48
##          949  954   985  1046  1090  1091  1207  1216  1287  1293
## 1312
## Predicted    11.90 187  4.54 168.4 -0.799 19.01 -2.89 -3.74 -3.02  0.511
## 8.59
## cvpred      12.23 161  8.01 134.1  2.212 18.74 -4.92 -20.27 -3.29  1.628
## 11.35
## num_deaths   14.00 143  0.00 172.0  0.000 20.00  0.00  0.00  0.00  0.000
## 0.00
## CV residual  1.77 -18 -8.01  37.9 -2.212  1.26  4.92  20.27  3.29 -1.628 -
## 11.35
##          1362 1428 1433 1456  1772  1815  1866  1876  1879 1929  1957
## 2010
## Predicted    -16.0 23.0 31.5 62.8  34.0  18.7  4.66  205  1.96 25.3  6.42
## 17.34
## cvpred      -18.2 18.1 10.0 66.2  38.1  30.7  15.48 260  4.46 26.0  7.43
## 29.06
## num_deaths   0.0 29.0 53.0 85.0  27.0  0.0  0.00 104  0.00 19.0  0.00
## 33.00
## CV residual 18.2 10.9 43.0 18.8 -11.1 -30.7 -15.48 -156 -4.46 -7.0 -7.43
## 3.94
##          2062 2106  2117     2122  2140  2195  2289
## Predicted    48.4 205   6.69 -1.07e-14  9.40  1.96  15.0
## cvpred      53.2 260  14.12 4.52e+01  8.56  4.46  13.9
## num_deaths  55.0 126  0.00 0.00e+00  0.00  0.00  0.0
## CV residual  1.8 -134 -14.12 -4.52e+01 -8.56 -4.46 -13.9
## 
## Sum of squares = 212632      Mean square = 4089      n = 52

```

```

## Overall (Sum over all 52 folds)
## ms
## 5428

```

Pruned *age45-54:cardiacArrest*

Adjusted R-squared: 0.254

Average MSE 5375

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionE <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` +
  alzheimer*`age45-54` + cardiacarrhythmia*`age45-54` + cerebrovascula*`age45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` + heartfailure*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-54` +
  malignant*`age45-54` +
  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*`age45-54` +
  renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + vascular*`age45-54` + Northeast*`age45-54` +
  South*`age45-54` + West*`age45-54` , data = train)
summary(model.interactionE)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##   influenza + injuryPoisoning + ischemicheart + malignant +
##   obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##   conditionrespiratoryarrest + sepsis + vascular + Northeast +
##   South + West + allothcondsandcauses * `age45-54` + alzheimer *
##   `age45-54` + cardiacarrhythmia * `age45-54` + cerebrovascula *
##   `age45-54` + chroniclr * `age45-54` + covid19 * `age45-54` +
##   diabetes * `age45-54` + heartfailure * `age45-54` + influenza *
##   `age45-54` + injuryPoisoning * `age45-54` + ischemicheart *
##   `age45-54` + malignant * `age45-54` + obesity * `age45-54` +
##   othercirculatorysystem * `age45-54` + otherrespsystem * `age45-54` +
##   renalfailure * `age45-54` + conditionrespiratoryarrest *

```

```

##      `age45-54` + vascular * `age45-54` + Northeast * `age45-54` +
##      South * `age45-54` + West * `age45-54`, data = train)
##
## Residuals:
##   Min    1Q Median    3Q   Max
## -195.9 -18.0  -4.4   4.3 981.0
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                41.46     12.55   3.30  0.00098
## ***
## `age45-54`                  16.61     12.59   1.32  0.18740
## adultrespdistSyn          -28.21     13.96  -2.02  0.04351
## *
## allothcondsandcauses       -1.22     16.30  -0.07  0.94054
## alzheimer                  -46.54     15.28  -3.05  0.00237
## **
## cardiacArrest              -39.72     14.16  -2.80  0.00513
## **
## cardiacarrhythmia         -45.34     17.16  -2.64  0.00835
## **
## cerebrovascula             -45.60     18.02  -2.53  0.01153
## *
## chroniclr                  -43.93     18.55  -2.37  0.01807
## *
## covid19                     18.94     16.18   1.17  0.24219
## diabetes                    -20.45     20.93  -0.98  0.32872
## heartfailure               -44.60     19.23  -2.32  0.02057
## *
## hypertensive                -30.63     13.72  -2.23  0.02578
## *
## influenza                   -1.57     16.96  -0.09  0.92603
## injuryPoisoning            -42.47     17.77  -2.39  0.01700
## *
## ischemicheart              -44.85     17.58  -2.55  0.01086
## *
## malignant                   -44.92     18.62  -2.41  0.01599
## *
## obesity                     -20.58     18.56  -1.11  0.26777
## othercirculatorysystem      -41.26     18.55  -2.22  0.02635
## *
## otherrespsystem             -43.63     19.24  -2.27  0.02353
## *
## renalfailure                -32.54     22.06  -1.47  0.14054
## conditionrespiratoryarrest -43.87     18.69  -2.35  0.01907
## *
## sepsis                      -37.60     13.45  -2.80  0.00527
## **
## vascular                    -46.30     15.38  -3.01  0.00267
## **

```

```

## Northeast                                4.37    9.76    0.45  0.65456
## South                                    9.21    8.60    1.07  0.28438
## West                                     4.81    8.61    0.56  0.57632
## `age45-54` :allothcondsandcauses      13.39   19.79    0.68  0.49882
## `age45-54` :alzheimer                  -31.24   18.25   -1.71  0.08727
.
## `age45-54` :cardiacarrhythmia        -26.90   21.89   -1.23  0.21930
## `age45-54` :cerebrovascula          -27.01   23.96   -1.13  0.25986
## `age45-54` :chroniclr                -21.16   22.42   -0.94  0.34557
## `age45-54` :covid19                  99.60   19.11    5.21  2.3e-07
*** 
## `age45-54` :diabetes                 -10.33   23.24   -0.44  0.65683
## `age45-54` :heartfailure            -30.14   28.65   -1.05  0.29298
## `age45-54` :influenza              53.63   20.61    2.60  0.00939
**
## `age45-54` :injuryPoisoning       -25.69   22.62   -1.14  0.25620
## `age45-54` :ischemicheart         -13.30   23.60   -0.56  0.57315
## `age45-54` :malignant              -23.75   25.20   -0.94  0.34617
## `age45-54` :obesity                -22.43   23.00   -0.98  0.32964
## `age45-54` :othercirculatorysystem -20.43   24.33   -0.84  0.40126
## `age45-54` :otherrespsystem       -27.83   25.67   -1.08  0.27867
## `age45-54` :renalfailure          -26.53   24.77   -1.07  0.28431
## `age45-54` :conditionrespiratoryarrest -32.51   24.45   -1.33  0.18379
## `age45-54` :vascular               -29.46   19.72   -1.49  0.13553
## `age45-54` :Northeast              24.87   13.45    1.85  0.06469
.
## `age45-54` :South                  18.19    11.86    1.53  0.12553
## `age45-54` :West                  17.96    12.47    1.44  0.14993
## --- 
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.2 on 1054 degrees of freedom
## Multiple R-squared:  0.286, Adjusted R-squared:  0.254
## F-statistic: 8.98 on 47 and 1054 DF, p-value: <2e-16

outE <- cv.lm(data = test
                , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West +
#
allothcondsandcauses*`age45-54` +
alzheimer*`age45-54` + cardiacarrhythmia*`age45-54` + cerebrovascula*`ag
e45-54` +
chroniclr*`age45-54` + covid19*`age45-54` + diabetes*`age45-54` + heartfa
ilure*`age45-54` +

```

```

influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-
54` + malignant*`age45-54` +
obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*
`age45-54` + renalfailure*`age45-54` +
conditionrespiratoryarrest*`age45-54` + vascular*`age45-54` + Northeast*
`age45-54` +
South*`age45-54` + West*`age45-54`)
, plotit = "Observed", m=5)

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Analysis of Variance Table
##
## Response: num_deaths
##                                     Df Sum Sq Mean Sq F value Pr(>F)
## `age45-54`                         1 55638 55638 11.60 0.00079
***  

## adultrespdistSyn                  1   308    308   0.06 0.80017
## allothcondsandcauses              1 124916 124916 26.04 7.4e-07
***  

## alzheimer                          1 17801 17801  3.71 0.05540
.  

## cardiacArrest                      1   9110   9110   1.90 0.16963
## cardiacarrhythmia                 1   1509   1509   0.31 0.57545
## cerebrovascula                     1    512    512   0.11 0.74420
## chroniclr                           1   1137   1137   0.24 0.62686
## covid19                             1 355946 355946 74.20 1.6e-15
***  

## diabetes                            1   3533   3533   0.74 0.39178
## heartfailure                        1    499    499   0.10 0.74730
## hypertensive                         1    196    196   0.04 0.83989
## influenza                            1   7406   7406   1.54 0.21542
## injuryPoisoning                     1    992    992   0.21 0.64974
## ischemicheart                        1    500    500   0.10 0.74723
## malignant                            1   1759   1759   0.37 0.54551
## obesity                              1     5     5   0.00 0.97305
## othercirculatorysystem               1     46    46   0.01 0.92211
## otherrespsystem                      1   203   203   0.04 0.83715
## renalfailure                         1   4668   4668   0.97 0.32504
## conditionrespiratoryarrest          1   1303   1303   0.27 0.60276

```

```

## sepsis                                1    176    176   0.04 0.84847
## vascular                               1 12311 12311  2.57 0.11065
## Northeast                             1    113    113   0.02 0.87800
## South                                  1   3229   3229   0.67 0.41286
## West                                   1   2523   2523   0.53 0.46916
## `age45-54` :allothcondsandcauses     1   82454   82454 17.19 4.9e-05
*** 

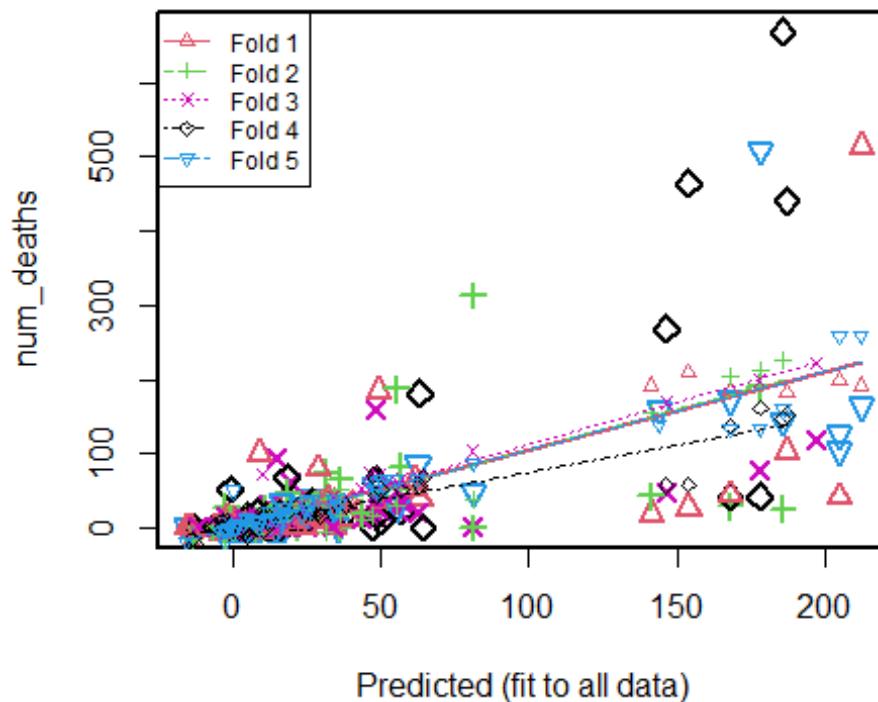
## `age45-54` :alzheimer                  1   1892   1892   0.39 0.53071
## `age45-54` :cardiacarrhythmia          1    626    626   0.13 0.71839
## `age45-54` :cerebrovascula             1    139    139   0.03 0.86518
## `age45-54` :chroniclr                 1    514    514   0.11 0.74386
## `age45-54` :covid19                   1   4531   4531   0.94 0.33223
## `age45-54` :diabetes                  1    498    498   0.10 0.74755
## `age45-54` :heartfailure              1     59     59   0.01 0.91160
## `age45-54` :influenza                1   550    550   0.11 0.73521
## `age45-54` :injuryPoisoning           1   373    373   0.08 0.78067
## `age45-54` :ischemicheart            1   1184   1184   0.25 0.61982
## `age45-54` :malignant                1   1323   1323   0.28 0.60005
## `age45-54` :obesity                  1    472    472   0.10 0.75405
## `age45-54` :othercirculatorysystem    1     16     16   0.00 0.95468
## `age45-54` :otherrespsystem          1   376    376   0.08 0.77981
## `age45-54` :renalfailure              1   1607   1607   0.33 0.56340
## `age45-54` :conditionrespiratoryarrest 1    578    578   0.12 0.72882
## `age45-54` :vascular                 1   4698   4698   0.98 0.32347
## `age45-54` :Northeast                 1   2246   2246   0.47 0.49454
## `age45-54` :South                    1     12     12   0.00 0.96005
## `age45-54` :West                     1   367    367   0.08 0.78230
## Residuals                            214 1026636   4797

## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
## :
## 
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

## Small symbols show cross-validation predicted values



```
## 
## fold 1
## Observations in test set: 52
##          21    27    38    79   186   280   359   396   413   41
## Predicted  0.0743 14.7  8.63  5.94  49.1  63.1  154  4.23 24.635 15.653
## cvpred    2.2969 16.6  4.41  9.80  33.7  51.4  209  9.98 30.994 20.982
## num_deaths 0.0000  0.0  0.00  0.00 185.0  39.0   26  0.00 30.000 21.000
## CV residual -2.2969 -16.6 -4.41 -9.80 151.3 -12.4 -183 -9.98 -0.994  0.017
##          499   509   544   578   609   668   675   699   726   728
## Predicted  33.9  1.95  30.3 -13.64 -2.22  8.78 -2.00 -4.29 187.1  30.3
## cvpred     44.1  3.60  40.7 -6.11 -1.41  5.89 -7.41 -8.65 182.7  40.7
## num_deaths  0.0  0.00  22.0  0.00  0.00 100.00 13.00  0.00 103.0  10.0
## CV residual -44.1 -3.60 -18.7  6.11  1.41  94.11 20.41  8.65 -79.7 -30.7
##          818   917   972  1007  1015  1100  1127 1139  1174  1266   1
## Predicted  205 24.63 10.83 12.81 -0.838  39.0  4.33 141 26.65 26.65   2
## cvpred     4.7
```

```

## cvpred      199 30.99 -6.92  6.96  0.418  35.5 13.00 192 21.97 21.97 2
7.6
## num_deaths  41 24.00  0.00 14.00 11.000 11.0   0.00   18 17.00 14.00
0.0
## CV residual -158 -6.99  6.92  7.04 10.582 -24.5 -13.00 -174 -4.97 -7.97 -2
7.6
##          1464 1500 1517 1564 1586 1589 1600 1689 1892 1909
## Predicted  40.36 -5.93 0.0557 22.3 61.4 -0.678 212 32.2 -14.3 4.33
## cvpred     38.91 -11.06 1.5312 12.7 50.1 -5.104 192 43.1 -16.1 13.00
## num_deaths 31.00  0.00 0.0000 14.0 64.0  0.000 514 10.0  0.0  0.00
## CV residual -7.91 11.06 -1.5312 1.3 13.9  5.104 322 -33.1 16.1 -13.00
##          1914 1965 1994 2071 2138 2174 2236 2275 2304 2313
## Predicted -16.03 32.2 28.9 7.93 11.54 21.85 168 11.8 21.10 1.75
## cvpred     -9.24 43.1 24.3 12.92 3.06 16.05 186 17.4 6.36 3.23
## num_deaths  0.00 39.0 79.0 0.00 0.00 15.00 44 0.0  0.00 0.00
## CV residual 9.24 -4.1 54.7 -12.92 -3.06 -1.05 -142 -17.4 -6.36 -3.23
##
## Sum of squares = 262496      Mean square = 5048      n = 52
##
## fold 2
## Observations in test set: 53
##          39 125 156 178 198 266 310 352 355 371
397
## Predicted  22.1 5.94 7.66 178.1 81.0 47.36 9.14 1.84 10.9 7.93
5.94
## cvpred     26.2 9.89 -7.85 211.9 37.2 15.92 17.46 2.52 17.1 11.09
9.89
## num_deaths  0.0 0.00 11.00 190.0 313.0 22.00 0.00 0.00 0.0  0.00
0.00
## CV residual -26.2 -9.89 18.85 -21.9 275.8 6.08 -17.46 -2.52 -17.1 -11.09
-9.89
##          412 457 458 492 503 517 571 586 635 656 736
## Predicted  56.9 25.7 56.88 55.1 15.6 -2.22 -2.0 168 31.08 -2.44 -2.90
## cvpred     51.3 30.1 51.31 29.4 18.9 -2.21 1.7 204 -4.82 -15.91 -6.15
## num_deaths 82.0 15.0 53.00 42.0 0.0 0.00 0.0 29 74.00 28.00 0.00
## CV residual 30.7 -15.1 1.69 12.6 -18.9 2.21 -1.7 -175 78.82 43.91 6.15
##          876 894 900 977 982 1003 1028 1034 1039 1041 1
084
## Predicted  23.0 12.3 11.5 0.339 36.2 31.08 11.00 36.2 -6.94 -1.04 -1
3.6
## cvpred     32.8 21.7 19.0 -2.258 51.7 -4.82 15.22 17.2 -6.08 -6.31 -1
7.5
## num_deaths 10.0 11.0 0.0 0.000 0.0 46.00 25.00 65.0 0.00 12.00
0.0
## CV residual -22.8 -10.7 -19.0 2.258 -51.7 50.82 9.78 47.8 6.08 18.31 1
7.5
##          1204 1222 1239 1288 1321 1511 1559 1567 1645 1675 1690
## Predicted  81.0 -3.54 24.6 -2.90 19.32 4.67 18.2 4.51 141 10.9 186
## cvpred     37.2 -9.46 28.5 -6.15 21.42 7.37 18.9 5.69 156 17.7 225
## num_deaths  0.0 0.00 14.0 0.00 14.00 0.00 47.0 12.00 43 0.0 25

```

```

## CV residual -37.2 9.46 -14.5 6.15 -7.42 -7.37 28.1 6.31 -113 -17.7 -200
##          1734 1773 1873 1983 2008 2035 2061 2097 2211
## Predicted 8.63 -0.678 32.20 13.33 55.1 -2.22 43.97 5.94 -4.29
## cvpred    8.44 -2.682 37.39 13.56 29.4 -2.21 11.15 9.89 -7.98
## num_deaths 0.00 0.000 42.00 17.00 187.0 0.00 20.00 0.00 0.00
## CV residual -8.44 2.682 4.61 3.44 157.6 2.21 8.85 -9.89 7.98
##
## Sum of squares = 209446      Mean square = 3952      n = 53
##
## fold 3
## Observations in test set: 53
##          16 190 209 211 274 307 325 428 454 487
## Predicted 81 7.20 0.781 14.72 55.4 -0.678 24.6 15.434 48.1 6.86
## cvpred    104 19.45 -9.365 -3.06 51.2 6.221 24.4 17.184 38.3 10.17
## num_deaths 0 14.00 13.000 93.00 23.0 0.000 0.0 17.000 158.0 0.00
## CV residual -104 -5.45 22.365 96.06 -28.2 -6.221 -24.4 -0.184 119.7 -10.17
##          491 527 597 677 774 793 868 904 907 909
967
## Predicted 11.85 14.72 0.0557 19.32 48.1 7.87 10.0 62.4 24.7 146
0.392
## cvpred    15.68 -3.06 -3.6346 6.92 38.3 14.54 72.1 69.3 22.5 169
5.851
## num_deaths 12.00 0.00 0.0000 61.00 18.0 0.00 10.0 21.0 40.0 48
0.000
## CV residual -3.68 3.06 3.6346 54.08 -20.3 -14.54 -62.1 -48.3 17.5 -121 -
5.851
##          994 1009 1014 1071 1213 1225 1327 1334 1375 1391
1438
## Predicted 9.14 11.75 -2.22 -2.71 -11.7 1.75 1.89 -2.9 18.2 0.339
22.7
## cvpred    12.13 4.86 -7.69 -2.34 -41.9 -1.51 2.58 -6.0 24.4 10.596
30.1
## num_deaths 0.00 32.00 15.00 0.00 0.0 0.00 0.00 0.0 0.0 0.000
13.0
## CV residual -12.13 27.14 22.69 2.34 41.9 1.51 -2.58 6.0 -24.4 -10.596
-17.1
##          1472 1523 1540 1554 1760 1805 1818 1836 1838 1936
1960
## Predicted 7.2 3.23 46.3 197 22.7 0.339 34.0 56.9 55.8 43.2
1.839
## cvpred    19.5 1.07 74.0 221 30.1 10.596 42.7 65.5 61.5 53.1 -
0.762
## num_deaths 0.0 0.00 15.0 119 11.0 0.000 0.0 33.0 49.0 13.0
0.000
## CV residual -19.5 -1.07 -59.0 -102 -19.1 -10.596 -42.7 -32.5 -12.5 -40.1
0.762
##          2041 2050 2052 2125 2152 2190 2225 2248 2252 225
5
## Predicted -11.7 -3.54 -5.93 2.64 49.1 178 4.33e+00 15.7 3.52 8.5
9

```

```

## cvpred      -41.9  6.85  5.59  13.24  71.8   198  1.78e-14  11.4   3.10  19.5
3
## num_deaths  0.0   0.00  0.00  0.00   25.0    78  1.30e+01   0.0   0.00  0.0
0
## CV residual 41.9  -6.85 -5.59 -13.24 -46.8  -120  1.30e+01 -11.4  -3.10 -19.5
3
##
## Sum of squares = 104246      Mean square = 1967      n = 53
##
## fold 4
## Observations in test set: 52
##          7    94   126   159   183   217   313   320   382   385
## Predicted 14.5  56.9  1.84 -0.0488  4.671 -4.154  3.34  48.1 15.43 10.05
## cvpred     17.4  50.3 -5.98 -2.4570 -0.256  0.856  2.57  43.7  9.76 -3.83
## num_deaths  0.0   23.0  0.00 10.0000 15.000  0.000  0.00  63.0  0.00  0.00
## CV residual -17.4 -27.3  5.98 12.4570 15.256 -0.856 -2.57 19.3 -9.76  3.83
##          434   550   628   630   647   680   761   808   881   936   944
## Predicted 29.63 37.96  47.4   8.63 -0.838  187  -6.03  4.23   50  28.87 26.6
## cvpred     15.56 30.35  70.6  17.45 -3.139  152   1.83  -4.75  51  20.65 20.3
## num_deaths 25.00 21.00   0.0   0.00 52.000  440   0.00   0.00 10  17.00 33.0
## CV residual  9.44 -9.35 -70.6 -17.45 55.139  288  -1.83  4.75  -41 -3.65 12.7
##          962   963   965   992  1083  1136  1138  1171  1197  1221   1
308
## Predicted 64.2  17.1   2.62  11.5  -6.03 -1.47 168.0 24.82 -2.17 -3.24 -1
2.9
## cvpred     67.4  15.3  -2.22  12.3   1.83 -7.28 136.9 17.65  1.62  1.49 -2
1.9
## num_deaths  0.0   0.0   0.00   0.0   0.00  0.00  42.0 12.00  0.00  0.00
0.0
## CV residual -67.4 -15.3   2.22 -12.3  -1.83  7.28 -94.9 -5.65 -1.62 -1.49
2
1.9
##          1315  1421  1461  1490  1506  1513  1544  1562  1626  1636  164
8
## Predicted -6.94 18.2 146.5  21.9   178   15.6 -3.54  63.1 11.00 -13.6  48.
1
## cvpred     1.20 25.1  57.4  30.1   162   27.2  9.78  57.6 10.46 -15.0  43.
7
## num_deaths  0.00 67.0 268.0  10.0    42   15.0  0.00 180.0 16.00   0.0  17.
0
## CV residual -1.20 41.9 210.6 -20.1  -120  -12.2 -9.78 122.4  5.54  15.0 -26.
7
##          1775  1776  1814  1860  1883  1979  2012  2013  2019
## Predicted -1.59  9.14 12.27  4.97  11.8  7.926 186 154.0 15.6
## cvpred     -1.25 11.07  5.34 -11.71  17.8  0.247 147  58.2 27.2
## num_deaths  0.00  0.00  0.00   0.00   0.0 22.000 668 464.0   0.0
## CV residual  1.25 -11.07 -5.34  11.71 -17.8 21.753 521 405.8 -27.2
##
## Sum of squares = 624865      Mean square = 12017      n = 52
##
## fold 5

```

```

## Observations in test set: 52
##      5    12    26    31   123   152   176   179   204   224    3
06
## Predicted  15.6  35.5 -2.8 -3.24  6.86  81.0  8.633 143.9 34.00 178  34
.0
## cvpred    12.5  33.2 -17.5 -3.04  8.61  86.4 -0.133 139.7 25.63 133  38
.1
## num_deaths  0.0   0.0   0.0   0.00  0.00  48.0 15.000 160.0 34.00 509  10
.0
## CV residual -12.5 -33.2  17.5  3.04 -8.61 -38.4 15.133 20.3  8.37 376 -28
.1
##          411   486   533   537   757   840   849   858   871   910
946
## Predicted  25.7  26.6 -4.15  0.781 11.9  15.43  24.8  55.1 17.10 212.3
4.23
## cvpred    24.2  38.0 -6.01 -0.479 22.0  22.11  35.0  65.1 16.91 259.3  1
4.40
## num_deaths 40.0  25.0  0.00  0.000  0.0 17.00  18.0  28.0 22.00 164.0
0.00
## CV residual 15.8 -13.0  6.01  0.479 -22.0 -5.11 -17.0 -37.1  5.09 -95.3 -1
4.40
##          949   954   985  1046  1090  1091  1207  1216  1287  1293
## Predicted 11.85 185.9  3.34 168.0 -1.47 19.319 -2.22 -2.8 -2.73  0.444
## cvpred   12.06 159.6  4.98 133.4  0.40 19.371 -3.42 -17.5 -2.68  1.358
## num_deaths 14.00 143.0  0.00 172.0  0.00 20.000  0.00  0.0  0.00  0.000
## CV residual 1.94 -16.6 -4.98  38.6 -0.40  0.629  3.42  17.5  2.68 -1.358
##          1312  1362  1428  1433  1456  1772  1815  1866  1876  1879  19
29
## Predicted  8.78 -16.0  23.0  50.0  62.4  34.0  17.3  4.23  205  1.89 24.
63
## cvpred   11.83 -18.3  18.1  63.9  65.2  38.1  27.4  14.40  259  4.12 24.
47
## num_deaths  0.00   0.0 29.0  53.0  85.0  27.0   0.0   0.00  104  0.00 19.
00
## CV residual -11.83  18.3 10.9 -10.9 19.8 -11.1 -27.4 -14.40 -155 -4.12 -5.
47
##          1957  2010  2062  2106   2117      2122  2140  2195  2289
## Predicted  6.86 16.40 48.13  205   4.51 -1.03e-13  9.14  1.89  14.5
## cvpred   8.61 26.58 52.62  259   10.74  5.04e+01  8.04  4.12  12.8
## num_deaths  0.00 33.00 55.00  126   0.00  0.00e+00  0.00  0.00  0.0
## CV residual -8.61  6.42  2.38 -133 -10.74 -5.04e+01 -8.04 -4.12 -12.8
##
## Sum of squares = 207243      Mean square = 3985      n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5375

##diabetes*age45-54 ## Adjusted R-squared: 0.255 ## MSE 5370

```

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionF <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` +
  alzheimer*`age45-54` + cardiacarrhythmia*`age45-54` + cerebrovascula*`age45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + heartfailure*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-54` +
  malignant*`age45-54` +
  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*`age45-54` +
  renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + vascular*`age45-54` + Northeast*`age45-54` +
  South*`age45-54` + West*`age45-54`, data = train)
summary(model.interactionF)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##   influenza + injuryPoisoning + ischemicheart + malignant +
##   obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##   conditionrespiratoryarrest + sepsis + vascular + Northeast +
##   South + West + allothcondsandcauses * `age45-54` + alzheimer *
##   `age45-54` + cardiacarrhythmia * `age45-54` + cerebrovascula *
##   `age45-54` + chroniclr * `age45-54` + covid19 * `age45-54` +
##   heartfailure * `age45-54` + influenza * `age45-54` + injuryPoisoning *
##   `age45-54` + ischemicheart * `age45-54` + malignant * `age45-54` +
##   obesity * `age45-54` + othercirculatorysystem * `age45-54` +
##   otherrespsystem * `age45-54` + renalfailure * `age45-54` +
##   conditionrespiratoryarrest * `age45-54` + vascular * `age45-54` +
##   Northeast * `age45-54` + South * `age45-54` + West * `age45-54`,
##   data = train)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -195.8   -18.2   -4.4    4.1   981.0
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   42.41     12.36   3.43  0.00063

```

| ***                                 |        |       |       |         |  |
|-------------------------------------|--------|-------|-------|---------|--|
| ## `age45-54`                       | 15.03  | 12.07 | 1.24  | 0.21352 |  |
| ## adultrespdistSyn                 | -28.28 | 13.95 | -2.03 | 0.04294 |  |
| *                                   |        |       |       |         |  |
| ## allothcondsandcauses             | -2.18  | 16.15 | -0.13 | 0.89283 |  |
| ## alzheimer                        | -47.49 | 15.12 | -3.14 | 0.00173 |  |
| **                                  |        |       |       |         |  |
| ## cardiacArrest                    | -39.81 | 14.16 | -2.81 | 0.00501 |  |
| **                                  |        |       |       |         |  |
| ## cardiacarrhythmia                | -46.28 | 17.02 | -2.72 | 0.00666 |  |
| **                                  |        |       |       |         |  |
| ## cerebrovascula                   | -46.54 | 17.89 | -2.60 | 0.00939 |  |
| **                                  |        |       |       |         |  |
| ## chroniclr                        | -44.89 | 18.42 | -2.44 | 0.01498 |  |
| *                                   |        |       |       |         |  |
| ## covid19                          | 17.98  | 16.03 | 1.12  | 0.26229 |  |
| ## diabetes                         | -27.55 | 13.51 | -2.04 | 0.04167 |  |
| *                                   |        |       |       |         |  |
| ## heartfailure                     | -45.55 | 19.11 | -2.38 | 0.01732 |  |
| *                                   |        |       |       |         |  |
| ## hypertensive                     | -30.62 | 13.71 | -2.23 | 0.02578 |  |
| *                                   |        |       |       |         |  |
| ## influenza                        | -2.54  | 16.81 | -0.15 | 0.88013 |  |
| ## injuryPoisoning                  | -43.42 | 17.63 | -2.46 | 0.01392 |  |
| *                                   |        |       |       |         |  |
| ## ischemicheart                    | -45.79 | 17.44 | -2.63 | 0.00878 |  |
| **                                  |        |       |       |         |  |
| ## malignant                        | -45.87 | 18.49 | -2.48 | 0.01326 |  |
| *                                   |        |       |       |         |  |
| ## obesity                          | -21.55 | 18.42 | -1.17 | 0.24248 |  |
| ## othercirculatorysystem           | -42.21 | 18.42 | -2.29 | 0.02211 |  |
| *                                   |        |       |       |         |  |
| ## otherrespystem                   | -44.57 | 19.12 | -2.33 | 0.01991 |  |
| *                                   |        |       |       |         |  |
| ## renalfailure                     | -33.50 | 21.95 | -1.53 | 0.12719 |  |
| ## conditionrespiratoryarrest       | -44.83 | 18.55 | -2.42 | 0.01584 |  |
| *                                   |        |       |       |         |  |
| ## sepsis                           | -37.74 | 13.44 | -2.81 | 0.00507 |  |
| **                                  |        |       |       |         |  |
| ## vascular                         | -47.27 | 15.22 | -3.10 | 0.00195 |  |
| **                                  |        |       |       |         |  |
| ## Northeast                        | 4.53   | 9.75  | 0.46  | 0.64270 |  |
| ## South                            | 9.25   | 8.60  | 1.08  | 0.28242 |  |
| ## West                             | 4.73   | 8.61  | 0.55  | 0.58294 |  |
| ## `age45-54` :allothcondsandcauses | 14.96  | 19.46 | 0.77  | 0.44214 |  |
| ## `age45-54` :alzheimer            | -29.67 | 17.90 | -1.66 | 0.09770 |  |
| .                                   |        |       |       |         |  |
| ## `age45-54` :cardiacarrhythmia    | -25.35 | 21.60 | -1.17 | 0.24074 |  |
| ## `age45-54` :cerebrovascula       | -25.46 | 23.70 | -1.07 | 0.28285 |  |
| ## `age45-54` :chroniclr            | -19.59 | 22.14 | -0.89 | 0.37636 |  |

```

## `age45-54` : covid19          101.16    18.78   5.39  8.8e-08
##
## `age45-54` : heartfailure     -28.60    28.43  -1.01  0.31455
## `age45-54` : influenza        55.21    20.30   2.72  0.00663
##
## `age45-54` : injuryPoisoning  -24.12    22.33  -1.08  0.28026
## `age45-54` : ischemicheart    -11.75    23.33  -0.50  0.61473
## `age45-54` : malignant         -22.19    24.95  -0.89  0.37383
## `age45-54` : obesity           -20.86    22.72  -0.92  0.35875
## `age45-54` : othercirculatorysystem -18.87    24.07  -0.78  0.43324
## `age45-54` : otherrespsystem   -26.29    25.43  -1.03  0.30151
## `age45-54` : renalfailure      -24.96    24.50  -1.02  0.30865
## `age45-54` : conditionrespiratoryarrest -30.94    24.18  -1.28  0.20093
## `age45-54` : vascular          -27.88    19.39  -1.44  0.15076
## `age45-54` : Northeast          24.74    13.44   1.84  0.06589
.
## `age45-54` : South              18.19    11.86   1.53  0.12533
## `age45-54` : West               18.11    12.46   1.45  0.14636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.2 on 1055 degrees of freedom
## Multiple R-squared:  0.286, Adjusted R-squared:  0.255
## F-statistic: 9.18 on 46 and 1055 DF, p-value: <2e-16

outF <- cv.lm(data = test
                , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
                  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
                  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
                  influenza + injuryPoisoning + ischemicheart + malignant +
                  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
                  conditionrespiratoryarrest + sepsis + vascular + Northeast +
                  Southwest +
                  ##
                  allothcondsandcauses*`age45-54` +
                  alzheimer*`age45-54` + cardiacarrhythmia*`age45-54` + cerebrovascula*`age45-54` +
                  chroniclr*`age45-54` + covid19*`age45-54` + heartfailure*`age45-54` +
                  influenza*`age45-54` + injuryPoisoning*`age45-54` + ischemicheart*`age45-54` +
                  malignant*`age45-54` +
                  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*`age45-54` +
                  renalfailure*`age45-54` +
                  conditionrespiratoryarrest*`age45-54` + vascular*`age45-54` + Northeast*`age45-54` +
                  South*`age45-54` + West*`age45-54`)
                , plotit = "Observed", m=5)

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction from a

```

```

## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Analysis of Variance Table
##
## Response: num_deaths
## Df Sum Sq Mean Sq F value Pr(>F)
## `age45-54` 1 55638 55638 11.65 0.00077
***  

## adultrespdistSyn 1 308 308 0.06 0.79972
## allothcondsandcauses 1 124916 124916 26.16 6.9e-07
***  

## alzheimer 1 17801 17801 3.73 0.05483
.  

## cardiacArrest 1 9110 9110 1.91 0.16863
## cardiacarrhythmia 1 1509 1509 0.32 0.57455
## cerebrovascula 1 512 512 0.11 0.74363
## chroniclr 1 1137 1137 0.24 0.62605
## covid19 1 355946 355946 74.54 1.3e-15
***  

## diabetes 1 3533 3533 0.74 0.39068
## heartfailure 1 499 499 0.10 0.74673
## hypertensive 1 196 196 0.04 0.83952
## influenza 1 7406 7406 1.55 0.21435
## injuryPoisoning 1 992 992 0.21 0.64898
## ischemicheart 1 500 500 0.10 0.74666
## malignant 1 1759 1759 0.37 0.54457
## obesity 1 5 5 0.00 0.97299
## othercirculatorysystem 1 46 46 0.01 0.92193
## otherrespsystem 1 203 203 0.04 0.83677
## renalfailure 1 4668 4668 0.98 0.32391
## conditionrespiratoryarrest 1 1303 1303 0.27 0.60191
## sepsis 1 176 176 0.04 0.84812
## vascular 1 12311 12311 2.58 0.10982
## Northeast 1 113 113 0.02 0.87771
## South 1 3229 3229 0.68 0.41177
## West 1 2523 2523 0.53 0.46812
## `age45-54`:allothcondsandcauses 1 82454 82454 17.27 4.7e-05
***  

## `age45-54`:alzheimer 1 1892 1892 0.40 0.52975
## `age45-54`:cardiacarrhythmia 1 626 626 0.13 0.71776
## `age45-54`:cerebrovascula 1 139 139 0.03 0.86487
## `age45-54`:chroniclr 1 514 514 0.11 0.74328

```

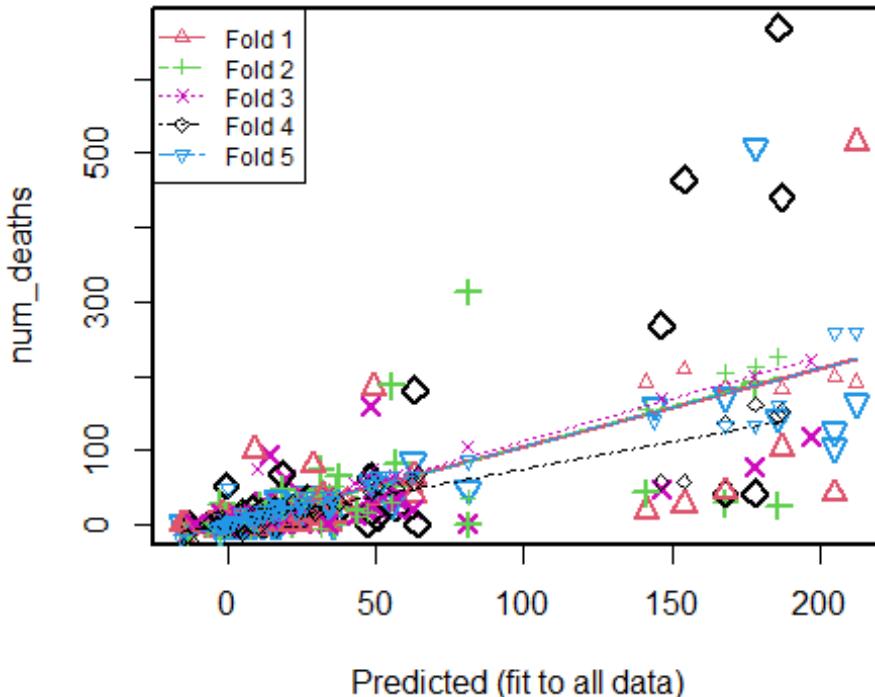
```

## `age45-54` : covid19           1   4531   4531  0.95 0.33110
## `age45-54` : heartfailure     1    84    84  0.02 0.89436
## `age45-54` : influenza        1   611   611  0.13 0.72092
## `age45-54` : injuryPoisoning  1   425   425  0.09 0.76571
## `age45-54` : ischemicheart    1  1255  1255  0.26 0.60872
## `age45-54` : malignant         1  1408  1408  0.29 0.58762
## `age45-54` : obesity           1   513   513  0.11 0.74336
## `age45-54` : othercirculatorysystem 1   23    23  0.00 0.94424
## `age45-54` : otherrespsystem   1   421   421  0.09 0.76670
## `age45-54` : renalfailure      1  1481  1481  0.31 0.57818
## `age45-54` : conditionrespiratoryarrest 1   663   663  0.14 0.70990
## `age45-54` : vascular          1  4835  4835  1.01 0.31542
## `age45-54` : Northeast          1  2251  2251  0.47 0.49310
## `age45-54` : South             1    13    13  0.00 0.95826
## `age45-54` : West              1   369   369  0.08 0.78119
## Residuals                      215 1026643  4775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
:
## 
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

## Small symbols show cross-validation predicted val



```

##  

## fold 1  

## Observations in test set: 52  

##  

##          21   27   38   79   186   280   359   396   413  

##  

## Predicted  0.0893 14.3  8.64  5.95  48.9  62.9  154  4.26 24.73 15.683  

##  

## cvpred    2.3027 16.3  4.41  9.81  33.6  51.3  209 10.00 31.04 21.000  

##  

## num_deaths 0.0000  0.0  0.00  0.00 185.0  39.0   26  0.00 30.00 21.000  

##  

## CV residual -2.3027 -16.3 -4.41 -9.81 151.4 -12.3 -183 -10.00 -1.04 -0.000  

##  

##          499   509   544   578   609   668   675   699   726   728  

##  

## Predicted 33.9  2.04  30.2 -13.67 -2.21   8.75 -2.00 -4.30 187.1  30.2  

## 1.87  

## cvpred    44.1  3.66  40.7 -6.15 -1.41   5.84 -7.41 -8.65 182.7  40.7  

## 6.87  

## num_deaths 0.0  0.00  22.0  0.00  0.00 100.00 13.00  0.00 103.0  10.0  

## 0.00  

## CV residual -44.1 -3.66 -18.7   6.15  1.41  94.16 20.41  8.65 -79.7 -30.7  

## -6.87  

##          818   917   972  1007  1015  1100  1127  1139  1174  1266   1  

## 413  

## Predicted 205 24.73 10.81 12.94 -0.749  38.8  4.33  141 26.68 26.68   2  

## 4.6  

## cvpred    199 31.04 -6.93  7.02  0.486  35.4  13.00 192 21.99 21.99   2  

## 7.5  

## num_deaths 41 24.00  0.00 14.00 11.000  11.0  0.00  18 17.00 14.00  

## 0.0  

## CV residual -158 -7.04  6.93  6.98 10.514 -24.4 -13.00 -174 -4.99 -7.99 -2  

## 7.5  

##          1464  1500  1517  1564  1586  1589  1600  1689  1892  1909  

## Predicted 40.36 -5.93  0.0669 22.28 62.5 -0.702  212  32.2 -14.4  4.33  

## cvpred    38.87 -11.08 1.5351 12.74 50.8 -5.122  192  43.1 -16.1 13.00  

## num_deaths 31.00  0.00  0.0000 14.00 64.0  0.000  514  10.0  0.0  0.00  

## CV residual -7.87 11.08 -1.5351  1.26 13.2  5.122  322 -33.1 16.1 -13.00  

##          1914  1965 1994  2071 2138  2174  2236  2275  2304  2313  

## Predicted -16.06 32.21 28.9  7.93 11.52 21.724  168 11.8 21.10  1.75  

## cvpred    -9.28 43.11 24.3 12.93  3.07 15.961  186 17.4  6.36  3.22  

## num_deaths 0.00 39.00 79.0  0.00  0.00 15.000   44  0.0  0.00  0.00  

## CV residual  9.28 -4.11 54.7 -12.93 -3.07 -0.961 -142 -17.4 -6.36 -3.22  

##  

##  

## Sum of squares = 262489      Mean square = 5048      n = 52  

##  

## fold 2  

## Observations in test set: 53  

##  

##          39   125   156   178   198   266   310   352   355   371  

## 397

```

```

## Predicted    22.1  5.95  7.66 178.1   80.9 47.36   9.12  1.87 10.9   7.93
5.95
## cvpred     26.3  9.96 -7.92 211.8   36.6 15.92  17.32  2.64 17.2 11.14
9.96
## num_deaths  0.0   0.00 11.00 190.0  313.0 22.00   0.00  0.00  0.0   0.00
0.00
## CV residual -26.3 -9.96 18.92 -21.8  276.4  6.08 -17.32 -2.64 -17.2 -11.14
-9.96
##          412   457   458   492   503   517   571   586   635   656   73
6
## Predicted   56.7  25.8 56.75 55.2   15.7 -2.21 -2.00  168 31.08 -2.47 -2.9
1
## cvpred      50.8  30.7 50.78 29.6   19.5 -2.17  1.71  204 -4.87 -15.83 -6.0
2
## num_deaths  82.0  15.0 53.00 42.0   0.0   0.00  0.00   29 74.00 28.00  0.0
0
## CV residual 31.2 -15.7 2.22 12.4  -19.5  2.17 -1.71 -175 78.87 43.83  6.0
2
##          876   894   900   977   982  1003  1028  1034  1039  1041   1
084
## Predicted   22.9  12.3 11.5  0.317  36.2 31.08 10.97 37.3 -6.94 -1.04 -1
3.7
## cvpred      32.7  21.5 18.8 -2.339  51.5 -4.87 15.24 23.2 -6.10 -6.39 -1
7.5
## num_deaths  10.0  11.0  0.0  0.000   0.0 46.00 25.00 65.0  0.00 12.00
0.0
## CV residual -22.7 -10.5 -18.8  2.339 -51.5 50.87  9.76 41.8  6.10 18.39
1
7.5
##          1204  1222  1239  1288  1321  1511 1559  1567 1645 1675 1690
## Predicted   80.9 -3.54 24.7 -2.91 19.32  4.68 18.3  4.57 141 10.9 186
## cvpred      36.6 -9.54 28.9 -6.02 21.41  7.41 19.3  6.00 156 17.8 225
## num_deaths  0.0   0.00 14.0  0.00 14.00  0.00 47.0 12.00 43  0.0 25
## CV residual -36.6  9.54 -14.9  6.02 -7.41 -7.41 27.7  6.00 -113 -17.8 -200
##          1734  1773  1873  1983  2008  2035 2061  2097 2211
## Predicted   8.64 -0.702 32.21 13.33  55.2 -2.21 43.97  5.95 -4.30
## cvpred      8.36 -2.778 37.48 13.63  29.6 -2.17 11.19  9.96 -8.03
## num_deaths  0.00  0.000 42.00 17.00 187.0  0.00 20.00  0.00  0.00
## CV residual -8.36  2.778  4.52  3.37 157.4  2.17  8.81 -9.96  8.03
##
## Sum of squares = 209208      Mean square = 3947      n = 53
##
## fold 3
## Observations in test set: 53
##          16   190   209   211   274   307   325   428   454
487
## Predicted   80.9  7.22  0.793 14.251  55.4 -0.702  24.7 15.460 48.2   6
.87
## cvpred     104.6 18.82 -9.272  0.904  51.5  6.386  23.6 17.136 38.2 10
.03
## num_deaths  0.0 14.00 13.000 93.000  23.0  0.000   0.0 17.000 158.0   0

```

```

.00
## CV residual -104.6 -4.82 22.272 92.096 -28.5 -6.386 -23.6 -0.136 119.8 -10
.03
##          491     527     597     677     774     793     868     904     907     909
## Predicted 11.85 14.251  0.0669 19.32   48.2    7.89   10.0    62.4   24.6   146
## cvpred    15.62 0.904 -3.6123  6.89   38.2   14.31   74.9   69.7   22.7   169
## num_deaths 12.00 0.000  0.0000 61.00   18.0    0.00   10.0   21.0   40.0    48
## CV residual -3.62 -0.904  3.6123 54.11  -20.2  -14.31  -64.9  -48.7  17.3  -121
##          967     994    1009   1014   1071   1213   1225   1327   1334   1375
## Predicted  0.362  9.12  11.83  -2.24  -2.7  -11.5   1.75   1.88  -2.91  18.3
## cvpred     6.105 12.34  4.09  -7.55  -2.5  -44.0  -1.36   2.50  -6.00  23.6
## num_deaths  0.000  0.00  32.00 15.00   0.0    0.0    0.00   0.00   0.00   0.0
## CV residual -6.105 -12.34 27.91 22.55   2.5   44.0   1.36  -2.50   6.00  -23.6
##          1391    1438   1472   1523   1540   1554   1760   1805   1818   183
6
## Predicted  0.317  22.7   7.22   3.23   47.5   197   22.7   0.317  33.9   56.
7
## cvpred     10.672 30.4   18.82   1.26   63.7   221   30.4   10.672 42.9   66.
8
## num_deaths  0.000  13.0   0.00   0.00   15.0   119   11.0   0.000  0.0   33.
0
## CV residual -10.672 -17.4  -18.82  -1.26  -48.7  -102  -19.4  -10.672 -42.9  -33.
8
##          1838    1936   1960   2041   2050   2052   2125   2152   2190   222
5
## Predicted  55.6   43.1   1.870  -11.5  -3.54  -5.93   2.62   48.9   178  4.33e+0
0
## cvpred     63.2   55.1  -0.967  -44.0   6.45   5.17  13.34  73.0   198  7.73e-1
4
## num_deaths  49.0   13.0   0.000   0.0   0.00   0.00   0.00   25.0   78  1.30e+0
1
## CV residual -14.2  -42.1   0.967  44.0  -6.45  -5.17  -13.34  -48.0  -120  1.30e+0
1
##          2248    2252   2255
## Predicted  15.7   3.56   8.59
## cvpred     11.1   2.68  19.48
## num_deaths  0.0   0.00   0.00
## CV residual -11.1  -2.68  -19.48
##
## Sum of squares = 103592      Mean square = 1955      n = 53
##
## fold 4
## Observations in test set: 52
##          7     94    126     159     183     217     313     320     382     385
## Predicted 14.6   56.7   1.87  -0.0489   4.676  -4.148   3.32   48.2  15.46 10.05
## cvpred    17.5   50.2  -5.95  -2.1664  -0.246   0.876   2.54   43.7   9.79  -3.56
## num_deaths 0.0   23.0   0.00 10.0000  15.000   0.000   0.00   63.0   0.00   0.00
## CV residual -17.5  -27.2   5.95 12.1664  15.246  -0.876  -2.54  19.3  -9.79  3.56
##          434    550    628    630    647    680    761    808    881    936    9

```

```

## Predicted 29.52 37.71 47.4 8.64 -0.749 187 -6.02 4.26 50.2 28.90 26
.7
## cvpred 15.39 30.11 70.6 17.46 -3.047 152 1.85 -4.73 51.2 20.68 20
.3
## num_deaths 25.00 21.00 0.0 0.00 52.000 440 0.00 0.00 10.0 17.00 33
.0
## CV residual 9.61 -9.11 -70.6 -17.46 55.047 288 -1.85 4.73 -41.2 -3.68 12
.7
## 962 963 965 992 1083 1136 1138 1171 1197 1221 1
308
## Predicted 64.0 17.2 2.59 11.5 -6.02 -1.49 168.0 24.35 -2.16 -3.23 -1
2.9
## cvpred 67.2 15.4 -2.26 12.3 1.85 -7.31 136.9 17.31 1.65 1.52 -2
1.9
## num_deaths 0.0 0.0 0.00 0.0 0.00 0.00 42.0 12.00 0.00 0.00
0.0
## CV residual -67.2 -15.4 2.26 -12.3 -1.85 7.31 -94.9 -5.31 -1.65 -1.52 2
1.9
## 1315 1421 1461 1490 1506 1513 1544 1562 1626 1636 164
8
## Predicted -6.94 18.3 146.4 21.7 178 15.7 -3.54 62.9 10.97 -13.7 48.
2
## cvpred 1.20 25.2 57.4 29.9 162 27.4 9.79 57.4 10.43 -15.0 43.
7
## num_deaths 0.00 67.0 268.0 10.0 42 15.0 0.00 180.0 16.00 0.0 17.
0
## CV residual -1.20 41.8 210.6 -19.9 -120 -12.4 -9.79 122.6 5.57 15.0 -26.
7
## 1775 1776 1814 1860 1883 1979 2012 2013 2019
## Predicted -1.62 9.12 12.26 5.0 11.8 7.934 186 154.0 15.7
## cvpred -1.27 11.05 5.32 -11.7 17.8 0.251 147 58.2 27.4
## num_deaths 0.00 0.00 0.00 0.0 0.0 22.000 668 464.0 0.0
## CV residual 1.27 -11.05 -5.32 11.7 -17.8 21.749 521 405.8 -27.4
##
## Sum of squares = 624888 Mean square = 12017 n = 52
##
## fold 5
## Observations in test set: 52
## 5 12 26 31 123 152 176 179 204 224 30
6
## Predicted 15.7 35.3 -2.8 -3.23 6.87 80.9 8.636 143.9 34.0 178 33.
9
## cvpred 13.6 31.8 -17.5 -2.65 8.75 86.0 -0.109 140.1 23.9 133 37.
9
## num_deaths 0.0 0.0 0.0 0.00 0.00 48.0 15.000 160.0 34.0 509 10.
0
## CV residual -13.6 -31.8 17.5 2.65 -8.75 -38.0 15.109 19.9 10.1 376 -27.
9
## 411 486 533 537 757 840 849 858 871 910
946

```

```

## Predicted 25.8 26.7 -4.15 0.7925 11.5 15.46 24.3 55.2 17.15 212.3
4.26
## cvpred 25.0 38.2 -5.70 -0.0624 17.8 22.33 30.9 65.2 17.23 259.3
14.66
## num_deaths 40.0 25.0 0.00 0.0000 0.0 17.00 18.0 28.0 22.00 164.0
0.00
## CV residual 15.0 -13.2 5.70 0.0624 -17.8 -5.33 -12.9 -37.2 4.77 -95.3 -
14.66
## 949 954 985 1046 1090 1091 1207 1216 1287 1293
## Predicted 11.85 185.9 3.32 168.0 -1.492 19.321 -2.21 -2.8 -2.73 0.438
## cvpred 11.99 159.8 4.66 133.3 0.327 19.275 -3.15 -17.5 -2.94 1.048
## num_deaths 14.00 143.0 0.00 172.0 0.000 20.000 0.00 0.0 0.00 0.000
## CV residual 2.01 -16.8 -4.66 38.7 -0.327 0.725 3.15 17.5 2.94 -1.048
## 1312 1362 1428 1433 1456 1772 1815 1866 1876 1879 19
29
## Predicted 8.75 -16.1 22.9 50.2 62.4 33.9 16.8 4.26 205 1.88 24.
73
## cvpred 11.60 -18.4 18.0 65.0 65.0 37.9 23.1 14.66 259 3.83 25.
08
## num_deaths 0.00 0.0 29.0 53.0 85.0 27.0 0.0 0.00 104 0.00 19.
00
## CV residual -11.60 18.4 11.0 -12.0 20.0 -10.9 -23.1 -14.66 -155 -3.83 -6.
08
## 1957 2010 2062 2106 2117 2122 2140 2195 2289
## Predicted 6.87 16.44 48.16 205 4.57 -8.53e-14 9.12 1.88 14.6
## cvpred 8.75 26.88 52.79 259 10.83 4.87e+01 7.91 3.83 13.7
## num_deaths 0.00 33.00 55.00 126 0.00 0.00e+00 0.00 0.00 0.0
## CV residual -8.75 6.12 2.21 -133 -10.83 -4.87e+01 -7.91 -3.83 -13.7
##
## Sum of squares = 206720 Mean square = 3975 n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5370

```

Pruned *age45-54:ischemicheart*

Adjusted R-squared: 0.255

MSE 5375

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionG <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothc
ondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` +

```

```

alzheimer*`age45-54` + cardiacarrhythmia*`age45-54` + cerebrovascula*`ag
e45-54` +
  chroniclr*`age45-54` + covid19*`age45-54` + heartfailure*`age45-54` +
  influenza*`age45-54` + injuryPoisoning*`age45-54` + malignant*`age45-54` +
+
  obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*`age45-54` +
  renalfailure*`age45-54` +
  conditionrespiratoryarrest*`age45-54` + vascular*`age45-54` + Northeast*`age45-54` +
  South*`age45-54` + West*`age45-54`, data = train)
summary(model.interactionG)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##      alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##      chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##      influenza + injuryPoisoning + ischemicheart + malignant +
##      obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##      conditionrespiratoryarrest + sepsis + vascular + Northeast +
##      South + West + allothcondsandcauses * `age45-54` + alzheimer *
##      `age45-54` + cardiacarrhythmia * `age45-54` + cerebrovascula *
##      `age45-54` + chroniclr * `age45-54` + covid19 * `age45-54` +
##      heartfailure * `age45-54` + influenza * `age45-54` + injuryPoisoning *
##      `age45-54` + malignant * `age45-54` + obesity * `age45-54` +
##      othercirculatorysystem * `age45-54` + otherrespsystem * `age45-54` +
##      renalfailure * `age45-54` + conditionrespiratoryarrest *
##      `age45-54` + vascular * `age45-54` + Northeast * `age45-54` +
##      South * `age45-54` + West * `age45-54`, data = train)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -195.8   -18.6   -4.3    4.8   981.0
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   43.42     12.20   3.56  0.00039
## ***
## `age45-54`                     13.38     11.62   1.15  0.24965
## adultrespdistSyn                -28.32     13.95  -2.03  0.04250
## *
## allothcondsandcauses            -3.08     16.05  -0.19  0.84776
## alzheimer                      -48.41     15.00  -3.23  0.00129
## **
## cardiacArrest                  -39.91     14.15  -2.82  0.00489
## **
## cardiacarrhythmia               -47.22     16.91  -2.79  0.00534
## **
## cerebrovascula                 -47.47     17.78  -2.67  0.00772

```

|                                           |        |       |       |         |  |
|-------------------------------------------|--------|-------|-------|---------|--|
| **                                        |        |       |       |         |  |
| ## chroniclr                              | -45.80 | 18.33 | -2.50 | 0.01259 |  |
| *                                         |        |       |       |         |  |
| ## covid19                                | 17.07  | 15.93 | 1.07  | 0.28397 |  |
| ## diabetes                               | -27.39 | 13.50 | -2.03 | 0.04270 |  |
| *                                         |        |       |       |         |  |
| ## heartfailure                           | -46.48 | 19.01 | -2.45 | 0.01464 |  |
| *                                         |        |       |       |         |  |
| ## hypertensive                           | -30.60 | 13.71 | -2.23 | 0.02583 |  |
| *                                         |        |       |       |         |  |
| ## influenza                              | -3.42  | 16.72 | -0.20 | 0.83772 |  |
| ## injuryPoisoning                        | -44.33 | 17.53 | -2.53 | 0.01158 |  |
| *                                         |        |       |       |         |  |
| ## ischemicheart                          | -50.86 | 14.23 | -3.57 | 0.00037 |  |
| ***                                       |        |       |       |         |  |
| ## malignant                              | -46.81 | 18.39 | -2.55 | 0.01103 |  |
| *                                         |        |       |       |         |  |
| ## obesity                                | -22.44 | 18.33 | -1.22 | 0.22113 |  |
| ## othercirculatorysystem                 | -43.12 | 18.32 | -2.35 | 0.01877 |  |
| *                                         |        |       |       |         |  |
| ## otherrespsystem                        | -45.52 | 19.02 | -2.39 | 0.01685 |  |
| *                                         |        |       |       |         |  |
| ## renalfailure                           | -34.43 | 21.86 | -1.57 | 0.11556 |  |
| ## conditionrespiratoryarrest             | -45.78 | 18.45 | -2.48 | 0.01326 |  |
| *                                         |        |       |       |         |  |
| ## sepsis                                 | -37.84 | 13.43 | -2.82 | 0.00494 |  |
| **                                        |        |       |       |         |  |
| ## vascular                               | -48.19 | 15.11 | -3.19 | 0.00147 |  |
| **                                        |        |       |       |         |  |
| ## Northeast                              | 4.38   | 9.74  | 0.45  | 0.65321 |  |
| ## South                                  | 9.05   | 8.58  | 1.05  | 0.29224 |  |
| ## West                                   | 4.71   | 8.60  | 0.55  | 0.58416 |  |
| ## `age45-54` :allothcondsandcauses       | 16.47  | 19.22 | 0.86  | 0.39179 |  |
| ## `age45-54` :alzheimer                  | -28.16 | 17.64 | -1.60 | 0.11073 |  |
| ## `age45-54` :cardiacarrhythmia          | -23.82 | 21.38 | -1.11 | 0.26537 |  |
| ## `age45-54` :cerebrovascula             | -23.94 | 23.49 | -1.02 | 0.30841 |  |
| ## `age45-54` :chroniclr                  | -18.08 | 21.92 | -0.82 | 0.40968 |  |
| ## `age45-54` :covid19                    | 102.67 | 18.53 | 5.54  | 3.8e-08 |  |
| ***                                       |        |       |       |         |  |
| ## `age45-54` :heartfailure               | -27.10 | 28.26 | -0.96 | 0.33787 |  |
| ## `age45-54` :influenza                  | 56.69  | 20.07 | 2.82  | 0.00483 |  |
| **                                        |        |       |       |         |  |
| ## `age45-54` :injuryPoisoning            | -22.61 | 22.12 | -1.02 | 0.30690 |  |
| ## `age45-54` :malignant                  | -20.64 | 24.75 | -0.83 | 0.40438 |  |
| ## `age45-54` :obesity                    | -19.37 | 22.52 | -0.86 | 0.38976 |  |
| ## `age45-54` :othercirculatorysystem     | -17.36 | 23.87 | -0.73 | 0.46722 |  |
| ## `age45-54` :otherrespsystem            | -24.75 | 25.24 | -0.98 | 0.32698 |  |
| ## `age45-54` :renalfailure               | -23.44 | 24.31 | -0.96 | 0.33519 |  |
| ## `age45-54` :conditionrespiratoryarrest | -29.40 | 23.98 | -1.23 | 0.22037 |  |
| ## `age45-54` :vascular                   | -26.37 | 19.15 | -1.38 | 0.16878 |  |

```

## `age45-54` :Northeast          24.89    13.43    1.85  0.06413
.
## `age45-54` :South            18.44    11.84    1.56  0.11980
## `age45-54` :West             18.21    12.45    1.46  0.14391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.1 on 1056 degrees of freedom
## Multiple R-squared:  0.286, Adjusted R-squared:  0.255
## F-statistic: 9.38 on 45 and 1056 DF,  p-value: <2e-16

outG <- cv.lm(data = test
                , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West +
##
allothcondsandcauses*`age45-54` +
alzheimer*`age45-54` + cardiacarrhythmia*`age45-54` + cerebrovascula*`ag
e45-54` +
chroniclr*`age45-54` + covid19*`age45-54` + heartfailure*`age45-54` +
influenza*`age45-54` + injuryPoisoning*`age45-54` + malignant*`age45-54` +
+
obesity*`age45-54` + othercirculatorysystem*`age45-54` + otherrespsystem*
`age45-54` + renalfailure*`age45-54` +
conditionrespiratoryarrest*`age45-54` + vascular*`age45-54` + Northeast*`ag
e45-54` +
South*`age45-54` + West*`age45-54`)
                , plotit = "Observed", m=5)

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Warning in predict.lm(subs.lm, newdata = data[rows.out, ]): prediction fro
m a
## rank-deficient fit may be misleading

## Analysis of Variance Table
##
## Response: num_deaths
##
```

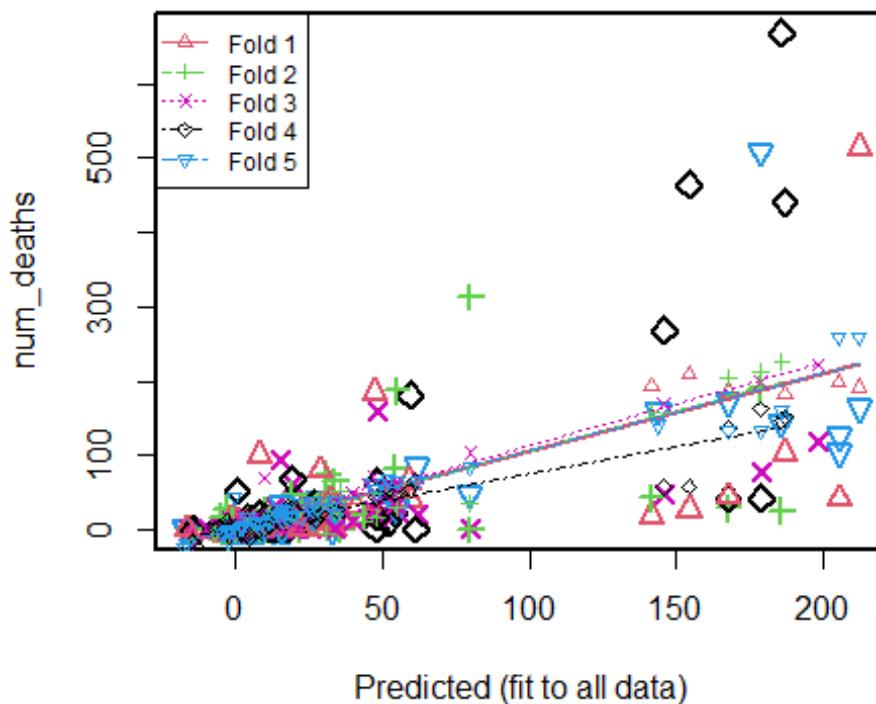
|                                           |   |        |        |       |         |
|-------------------------------------------|---|--------|--------|-------|---------|
| ## `age45-54`                             | 1 | 55638  | 55638  | 11.68 | 0.00075 |
| ***                                       |   |        |        |       |         |
| ## adultrespdistSyn                       | 1 | 308    | 308    | 0.06  | 0.79947 |
| ## allothcondsandcauses                   | 1 | 124916 | 124916 | 26.23 | 6.7e-07 |
| ***                                       |   |        |        |       |         |
| ## alzheimer                              | 1 | 17801  | 17801  | 3.74  | 0.05452 |
| .                                         |   |        |        |       |         |
| ## cardiacArrest                          | 1 | 9110   | 9110   | 1.91  | 0.16810 |
| ## cardiacarrhythmia                      | 1 | 1509   | 1509   | 0.32  | 0.57407 |
| ## cerebrovascula                         | 1 | 512    | 512    | 0.11  | 0.74331 |
| ## chroniclr                              | 1 | 1137   | 1137   | 0.24  | 0.62562 |
| ## covid19                                | 1 | 355946 | 355946 | 74.73 | 1.2e-15 |
| ***                                       |   |        |        |       |         |
| ## diabetes                               | 1 | 3533   | 3533   | 0.74  | 0.39008 |
| ## heartfailure                           | 1 | 499    | 499    | 0.10  | 0.74642 |
| ## hypertensive                           | 1 | 196    | 196    | 0.04  | 0.83933 |
| ## influenza                              | 1 | 7406   | 7406   | 1.55  | 0.21377 |
| ## injuryPoisoning                        | 1 | 992    | 992    | 0.21  | 0.64857 |
| ## ischemicheart                          | 1 | 500    | 500    | 0.10  | 0.74636 |
| ## malignant                              | 1 | 1759   | 1759   | 0.37  | 0.54406 |
| ## obesity                                | 1 | 5      | 5      | 0.00  | 0.97295 |
| ## othercirculatorysystem                 | 1 | 46     | 46     | 0.01  | 0.92183 |
| ## otherrespsystem                        | 1 | 203    | 203    | 0.04  | 0.83657 |
| ## renalfailure                           | 1 | 4668   | 4668   | 0.98  | 0.32330 |
| ## conditionrespiratoryarrest             | 1 | 1303   | 1303   | 0.27  | 0.60146 |
| ## sepsis                                 | 1 | 176    | 176    | 0.04  | 0.84793 |
| ## vascular                               | 1 | 12311  | 12311  | 2.58  | 0.10937 |
| ## Northeast                              | 1 | 113    | 113    | 0.02  | 0.87756 |
| ## South                                  | 1 | 3229   | 3229   | 0.68  | 0.41118 |
| ## West                                   | 1 | 2523   | 2523   | 0.53  | 0.46756 |
| ## `age45-54` :allothcondsandcauses       | 1 | 82454  | 82454  | 17.31 | 4.6e-05 |
| ***                                       |   |        |        |       |         |
| ## `age45-54` :alzheimer                  | 1 | 1892   | 1892   | 0.40  | 0.52924 |
| ## `age45-54` :cardiacarrhythmia          | 1 | 626    | 626    | 0.13  | 0.71742 |
| ## `age45-54` :cerebrovascula             | 1 | 139    | 139    | 0.03  | 0.86470 |
| ## `age45-54` :chroniclr                  | 1 | 514    | 514    | 0.11  | 0.74297 |
| ## `age45-54` :covid19                    | 1 | 4531   | 4531   | 0.95  | 0.33049 |
| ## `age45-54` :heartfailure               | 1 | 84     | 84     | 0.02  | 0.89423 |
| ## `age45-54` :influenza                  | 1 | 611    | 611    | 0.13  | 0.72058 |
| ## `age45-54` :injuryPoisoning            | 1 | 425    | 425    | 0.09  | 0.76542 |
| ## `age45-54` :malignant                  | 1 | 1256   | 1256   | 0.26  | 0.60816 |
| ## `age45-54` :obesity                    | 1 | 409    | 409    | 0.09  | 0.76973 |
| ## `age45-54` :othercirculatorysystem     | 1 | 7      | 7      | 0.00  | 0.96921 |
| ## `age45-54` :otherrespsystem            | 1 | 305    | 305    | 0.06  | 0.80054 |
| ## `age45-54` :renalfailure               | 1 | 1788   | 1788   | 0.38  | 0.54073 |
| ## `age45-54` :conditionrespiratoryarrest | 1 | 495    | 495    | 0.10  | 0.74753 |
| ## `age45-54` :vascular                   | 1 | 3818   | 3818   | 0.80  | 0.37161 |
| ## `age45-54` :Northeast                  | 1 | 2409   | 2409   | 0.51  | 0.47773 |
| ## `age45-54` :South                      | 1 | 7      | 7      | 0.00  | 0.96926 |
| ## `age45-54` :West                       | 1 | 542    | 542    | 0.11  | 0.73621 |

```

## Residuals                               216 1028840     4763
## ---
## Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
## :
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

## Small symbols show cross-validation predicted values



```

##
## fold 1
## Observations in test set: 52
##          21    27    38    79   186   280   359   396   413   418
499
## Predicted      0.143  15.5  9.04  6.06  47.3  59.5  154  4.21 26.10 15.837
33.6
## cvpred        2.353  16.9  4.64  9.91  32.6  49.3  209  9.34 31.95 20.405
43.7
## num_deaths    0.000    0.0    0.00    0.00 185.0   39.0    26    0.00 30.00 21.000
0.0
## CV residual -2.353 -16.9 -4.64 -9.91 152.4 -10.3 -183 -9.34 -1.95  0.595 -
43.7
##          509   544   578   609   668   675   699   726   728   810    8
18

```

```

## Predicted    3.20 29.9 -14.00 -2.00    8.60 -1.54 -4.22 187 29.9 1.87 2
05
## cvpred     4.94 41.5 -5.88 -1.21    6.25 -6.89 -8.56 183 41.5 6.32 1
98
## num_deaths  0.00 22.0  0.00  0.00 100.00 13.00  0.00 103 10.0  0.00
41
## CV residual -4.94 -19.5   5.88  1.21  93.75 19.89  8.56 -80 -31.5 -6.32 -1
57
##             917  972 1007 1015 1100 1127 1139 1174 1266 1413
1464
## Predicted   26.10 28.8 15.32 0.897 36.2  4.33 142 26.82 26.82 23.9 4
1.10
## cvpred     31.95 23.6  8.76  2.187 34.2  13.00 192 21.47 21.47 26.9 3
9.84
## num_deaths  24.00  0.0 14.00 11.000 11.0  0.00 18 17.00 14.00  0.0 3
1.00
## CV residual -7.95 -23.6   5.24  8.813 -23.2 -13.00 -174 -4.47 -7.47 -26.9 -8.84
##             1500 1517 1564 1586 1589 1600 1689 1892 1909 1914
1965
## Predicted   -5.19 0.107 21.98 59.2 -1.33 212 32.4 3.52 4.33 -16.3
32.36
## cvpred     -10.54 1.569 12.47 48.6 -5.69 192 43.2 14.56 13.00 -8.9
43.17
## num_deaths  0.00 0.000 14.00 64.0  0.00 514 10.0 0.00 0.00 0.00
39.00
## CV residual 10.54 -1.569  1.53 15.4  5.69 322 -33.2 -14.56 -13.00  8.9
-4.17
##             1994 2071 2138 2174 2236 2275 2304 2313
## Predicted   28.9 8.17 11.23 20.4636 168 11.9 21.9 1.54
## cvpred     23.7 13.17 3.14 15.0596 186 17.4 6.8 3.01
## num_deaths  79.0 0.00 0.00 15.0000 44 0.0 0.0 0.00
## CV residual 55.3 -13.17 -3.14 -0.0596 -142 -17.4 -6.8 -3.01
##
## Sum of squares = 263299      Mean square = 5063      n = 52
##
## fold 2
## Observations in test set: 53
##             39 125 156 178 198 266 310 352 355 371
397
## Predicted   22.0 6.06 8.40 178.7 79.9 47.95 8.89 1.87 11.1 8.17
6.06
## cvpred     26.5 9.57 -7.19 212.3 35.6 16.48 17.18 2.57 16.7 11.03
9.57
## num_deaths  0.0 0.00 11.00 190.0 313.0 22.00 0.00 0.00 0.0 0.00
0.00
## CV residual -26.5 -9.57 18.19 -22.3 277.4 5.52 -17.18 -2.57 -16.7 -11.03
-9.57
##             412 457 458 492 503 517 571 586 635 656 73
6

```

```

## Predicted 54.4 28.0 54.38 55.0 17.6 -2.00 -1.54 168 31.26 -2.76 -3.2
6
## cvpred 49.0 31.7 48.97 29.4 21.1 -1.82 2.52 203 -4.37 -16.05 -6.5
3
## num_deaths 82.0 15.0 53.00 42.0 0.0 0.00 0.00 29 74.00 28.00 0.0
0
## CV residual 33.0 -16.7 4.03 12.6 -21.1 1.82 -2.52 -174 78.37 44.05 6.5
3
## 876 894 900 977 982 1003 1028 1034 1039 1041 1
084
## Predicted 22.9 11.9 11.2 -0.122 35.9 31.26 10.7 34.0 -6.60 -0.764 -1
4.0
## cvpred 32.7 21.2 18.6 -2.605 51.3 -4.37 14.9 20.4 -5.47 -5.575 -1
7.8
## num_deaths 10.0 11.0 0.0 0.000 0.0 46.00 25.0 65.0 0.00 12.000
0.0
## CV residual -22.7 -10.2 -18.6 2.605 -51.3 50.37 10.1 44.6 5.47 17.575 1
7.8
## 1204 1222 1239 1288 1321 1511 1559 1567 1645 1675 1690
## Predicted 79.9 -2.85 26.1 -3.26 19.71 4.38 19.5 5.08 142 11.1 186
## cvpred 35.6 -8.92 29.6 -6.53 22.12 7.03 20.3 6.33 157 17.6 225
## num_deaths 0.0 0.00 14.0 0.00 14.00 0.00 47.0 12.00 43 0.0 25
## CV residual -35.6 8.92 -15.6 6.53 -8.12 -7.03 26.7 5.67 -114 -17.6 -200
## 1734 1773 1873 1983 2008 2035 2061 2097 2211
## Predicted 9.04 -1.33 32.36 -4.66 55.0 -2.00 43.92 6.06 -4.22
## cvpred 8.79 -3.20 37.17 -3.75 29.4 -1.82 10.67 9.57 -7.52
## num_deaths 0.00 0.00 42.00 17.00 187.0 0.00 20.00 0.00 0.00
## CV residual -8.79 3.20 4.83 20.75 157.6 1.82 9.33 -9.57 7.52
##
## Sum of squares = 210383 Mean square = 3969 n = 53
##
## fold 3
## Observations in test set: 53
## 16 190 209 211 274 307 325 428 454 48
7
## Predicted 79.9 7.9 0.769 15.47 55.2 -1.33 26.1 15.464 48.2 7.1
5
## cvpred 103.1 20.1 -10.026 1.96 50.9 5.44 25.6 17.239 38.4 10.5
7
## num_deaths 0.0 14.0 13.000 93.00 23.0 0.00 0.0 17.000 158.0 0.0
0
## CV residual -103.1 -6.1 23.026 91.04 -27.9 -5.44 -25.6 -0.239 119.6 -10.5
7
## 491 527 597 677 774 793 868 904 907 909
967
## Predicted 11.89 15.47 0.107 19.71 48.2 8.36 10.0 62.0 23.9 146 -0
.311
## cvpred 15.79 1.96 -3.996 7.56 38.4 15.50 68.8 68.9 21.5 168 4
.891
## num_deaths 12.00 0.00 0.000 61.00 18.0 0.00 10.0 21.0 40.0 48 0

```

```

.000
## CV residual -3.79 -1.96  3.996 53.44 -20.4 -15.50 -58.8 -47.9 18.5 -120 -4
.891
##          994 1009 1014 1071 1213 1225 1327 1334 1375   13
91
## Predicted    8.89 13.45 -2.38 -2.162 -9.23  1.54  2.08 -3.26 19.5 -0.1
22
## cvpred      11.92  6.53 -7.84 -0.973 -39.81 -2.30  3.26 -6.00 24.9 10.3
68
## num_deaths   0.00 32.00 15.00  0.000   0.00  0.00  0.00  0.00   0.0   0.0
00
## CV residual -11.92 25.47 22.84  0.973  39.81  2.30 -3.26 6.00 -24.9 -10.3
68
##          1438 1472 1523 1540 1554 1760   1805 1818 1836 1838
1936
## Predicted    22.5   7.9 -15.0  45.2  198  22.5 -0.122 33.8 54.4 52.5
40.0
## cvpred      29.7  20.1 -24.3  60.6  222  29.7 10.368 42.5 64.0 59.6
50.8
## num_deaths   13.0   0.0   0.0  15.0  119  11.0  0.000   0.0  33.0 49.0
13.0
## CV residual -16.7 -20.1  24.3 -45.6 -103 -18.7 -10.368 -42.5 -31.0 -10.6 -
37.8
##          1960 2041 2050 2052 2125 2152 2190   2225 2248   22
52
## Predicted    1.874 -9.23 -2.85 -5.19   2.02  47.3  179 4.33e+00 15.8 21
.7
## cvpred      -0.546 -39.81  7.24  6.02 12.96  70.6  199 4.62e-14 11.7 27
.8
## num_deaths   0.000   0.00  0.00  0.00   0.00  25.0   78 1.30e+01   0.0   0
.0
## CV residual  0.546  39.81 -7.24 -6.02 -12.96 -45.6 -121 1.30e+01 -11.7 -27
.8
##          2255
## Predicted    8.44
## cvpred      19.07
## num_deaths   0.00
## CV residual -19.07
##
## Sum of squares = 101792     Mean square = 1921      n = 53
##
## fold 4
## Observations in test set: 52
##          7   94  126   159   183   217   313   320   382   385
434
## Predicted    15.8  54.4  1.87 -0.176  4.383 -4.297  2.65 48.2 15.46 10.176
27.5
## cvpred      18.4  47.8 -5.94  2.258 -0.579  0.718  1.59 43.7  9.82  0.988
12.2
## num_deaths   0.0  23.0  0.00 10.000 15.000  0.000   0.00 63.0   0.00  0.000

```

```

25.0
## CV residual -18.4 -24.8  5.94  7.742 15.579 -0.718 -1.59 19.3 -9.82 -0.988
12.8
##          550   628    630    647  680    761   808   881   936   944    96
2
## Predicted 34.3  48.0   9.04  0.897 187 -5.51  4.21  52.1 28.91 26.8  61.
4
## cvpred    26.2  71.4  17.84 -1.071 151  2.56 -4.75  53.5 20.66 20.5  64.
5
## num_deaths 21.0   0.0   0.00 52.000 440   0.00   0.00 10.0 17.00 33.0   0.
0
## CV residual -5.2 -71.4 -17.84 53.071 289 -2.56  4.75 -43.5 -3.66 12.5 -64.
5
##          963   965   992  1083 1136  1138  1171  1197 1221 1308    1
315
## Predicted 17.6  1.98  11.2 -5.51 -2.12 167.5 25.83 -2.19 -3.21 -13.3 -6
.60
## cvpred    15.5 -3.17  11.9  2.56 -8.06 136.4 18.93  1.82  1.79 -22.7  1
.49
## num_deaths  0.0  0.00   0.0  0.00  0.00  42.0 12.00  0.00  0.00   0.0  0
.00
## CV residual -15.5  3.17 -11.9 -2.56  8.06 -94.4 -6.93 -1.82 -1.79 22.7 -1
.49
##          1421  1461  1490 1506 1513  1544  1562 1626 1636 1648    17
75
## Predicted 19.5 145.8  20.5  179  17.6 -2.85 59.5 10.69 -14.0 48.2 -2.
42
## cvpred    26.5  56.6  27.5  162  29.4 10.55 53.5 10.06 -15.3 43.7 -2.
19
## num_deaths 67.0 268.0 10.0   42  15.0   0.00 180.0 16.00   0.0 17.0  0.
00
## CV residual 40.5 211.4 -17.5 -120 -14.4 -10.55 126.5  5.94  15.3 -26.7  2.
19
##          1776  1814  1860 1883 1979 2012 2013 2019
## Predicted  8.89 11.90  4.88 13.4  8.166 186 154.2 17.6
## cvpred     10.73  4.58 -12.09 19.2  0.546 147  58.3 29.4
## num_deaths  0.00  0.00   0.00  0.0 22.000 668 464.0  0.0
## CV residual -10.73 -4.58 12.09 -19.2 21.454  521 405.7 -29.4
##
## Sum of squares = 626225      Mean square = 12043      n = 52
##
## fold 5
## Observations in test set: 52
##          5    12    26    31   123   152    176    179    204   224    3
06
## Predicted 17.6  32.9 -2.18 -3.21  7.15  79.9  9.036 143.9 34.0 179  33
.8
## cvpred    16.1  29.2 -16.63 -2.16  9.61  85.2  0.138 140.5 20.4 133  37
.8
## num_deaths  0.0   0.0   0.00  0.00  0.00  48.0 15.000 160.0 34.0 509  10

```

```

.0
## CV residual -16.1 -29.2 16.63 2.16 -9.61 -37.2 14.862 19.5 13.6 376 -27
.8
##          411 486 533 537 757 840 849 858 871 910 946
## Predicted 28.0 26.8 -4.30 0.769 13.2 15.46 25.8 55 17.63 212 4.21
## cvpred    27.9 38.3 -5.63 0.299 19.1 22.38 33.2 65 18.03 259 14.66
## num_deaths 40.0 25.0 0.00 0.000 0.0 17.00 18.0 28 22.00 164 0.00
## CV residual 12.1 -13.3 5.63 -0.299 -19.1 -5.38 -15.2 -37 3.97 -95 -14.66
##          949 954 985 1046 1090 1091 1207 1216 1287 1293
1312
## Predicted 11.89 185.7 2.65 167.5 -2.121 19.71 -2.00 -2.18 -2.20 -17.3
8.6
## cvpred     12.33 159.7 3.86 132.8 -0.413 18.92 -2.83 -16.63 -3.22 -23.9
11.4
## num_deaths 14.00 143.0 0.00 172.0 0.000 20.00 0.00 0.00 0.00 0.00 0.0
0.0
## CV residual 1.67 -16.7 -3.86 39.2 0.413 1.08 2.83 16.63 3.22 23.9
-11.4
##          1362 1428 1433 1456 1772 1815 1866 1876 1879 1929 195
7
## Predicted -16.3 22.9 52.1 62.0 33.8 17.3 4.21 205 2.08 26.10 7.1
5
## cvpred     -18.8 17.9 67.4 64.5 37.8 24.1 14.66 260 3.44 27.12 9.6
1
## num_deaths 0.0 29.0 53.0 85.0 27.0 0.0 0.00 104 0.00 19.00 0.0
0
## CV residual 18.8 11.1 -14.4 20.5 -10.8 -24.1 -14.66 -156 -3.44 -8.12 -9.6
1
##          2010 2062 2106 2117 2122 2140 2195 2289
## Predicted 16.10 48.16 205 5.08 -9.24e-14 8.89 2.08 15.8
## cvpred    26.49 52.83 260 11.29 4.40e+01 7.61 3.44 15.4
## num_deaths 33.00 55.00 126 0.00 0.00e+00 0.00 0.00 0.0
## CV residual 6.51 2.17 -134 -11.29 -4.40e+01 -7.61 -3.44 -15.4
##
## Sum of squares = 206600      Mean square = 3973      n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5375

```

Prunned *age45-54:chroniclr* , *age45-54:obesity* , *age45-54:heartfailure* , *age45-54:cerebrovascula* , ##*age45-54:otherrespsystem*,*age45-54:malignant* , *age45-54:injuryPoisoning* , *age45-54:cardiacarrhythmia* , ##*age45-54:conditionrespiratoryarrest* , *age45-54:vascular* , age45-54:*alzheimer* , *age45-54:renalfailure* , *age45-54:othercirculatorysystem*  
*covidcleansed\$`age45-54SQ`* <- *covidcleansed\$`age45-54`^2*  
*model.interactionH* <- *lm(num\_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +*

```

alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West +
##
allothcondsandcauses*`age45-54` + covid19*`age45-54` + influenza*`age
45-54` + Northeast*`age45-54` + South*`age45-54` + West*`age45-54`, data =
train)
summary(model.interactionH)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##      alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##      chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##      influenza + injuryPoisoning + ischemicheart + malignant +
##      obesity + othercirculatorysystem + otherrespsystem + renalfailure +
##      conditionrespiratoryarrest + sepsis + vascular + Northeast +
##      South + West + allothcondsandcauses * `age45-54` + covid19 *
##      `age45-54` + influenza * `age45-54` + Northeast * `age45-54` +
##      South * `age45-54` + West * `age45-54`, data = train)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -195.8   -20.2   -4.1    6.4   980.2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  53.70     11.42     4.70  2.9e-06 ***
## `age45-54`                   -2.90      9.50    -0.31  0.76006
## adultrespdistSyn             -28.89    13.90    -2.08  0.03792 *
## allothcondsandcauses        -12.23    15.55    -0.79  0.43186
## alzheimer                    -63.69    12.14    -5.25  1.9e-07 ***
## cardiacArrest                 -40.98    14.10    -2.91  0.00373 **
## cardiacarrhythmia            -60.32    13.53    -4.46  9.1e-06 ***
## cerebrovascula                -60.37    14.33    -4.21  2.7e-05 ***
## chroniclr                     -56.51    13.74    -4.11  4.2e-05 ***
## covid19                        7.83     15.42     0.51  0.61186
## diabetes                      -26.03    13.45    -1.94  0.05316 .
## heartfailure                  -59.83    15.86    -3.77  0.00017 ***
## hypertensive                   -30.45    13.66    -2.23  0.02604 *
## influenza                      -12.44    16.24    -0.77  0.44404
## injuryPoisoning               -56.95    13.85    -4.11  4.2e-05 ***
## ischemicheart                  -54.07    14.13    -3.83  0.00014 ***
## malignant                      -58.39    14.82    -3.94  8.7e-05 ***
## obesity                        -33.67    14.00    -2.40  0.01637 *
## othercirculatorysystem         -53.27    14.52    -3.67  0.00026 ***

```

```

## otherrespstystem          -58.90    15.06   -3.91  9.8e-05 ***
## renalfailure              -49.30    14.01   -3.52  0.00045 ***
## conditionrespiratoryarrest -61.49    14.58   -4.22  2.7e-05 ***
## sepsis                     -39.01    13.38   -2.92  0.00362 **
## vascular                  -61.79    12.58   -4.91  1.0e-06 ***
## Northeast                  2.98     9.66    0.31   0.75785
## South                      7.48     8.50    0.88   0.37889
## West                       3.71     8.54    0.43   0.66432
## `age45-54` :allothcondsandcauses 31.46    18.14   1.73   0.08326 .
## `age45-54` :covid19         117.89   17.38   6.78   2.0e-11 ***
## `age45-54` :influenza      71.64    19.04   3.76   0.00018 ***
## `age45-54` :Northeast       26.26    13.34   1.97   0.04929 *
## `age45-54` :South           20.81    11.73   1.77   0.07626 .
## `age45-54` :West            18.07    12.35   1.46   0.14385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.9 on 1069 degrees of freedom
## Multiple R-squared:  0.281, Adjusted R-squared:  0.26
## F-statistic: 13.1 on 32 and 1069 DF, p-value: <2e-16

outH <- cv.lm(data = test
               , form.lm = formula(num_deaths ~ `age45-54` + adultrespdistSyn +
allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespstystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West +
#
allothcondsandcauses*`age45-54` + covid19*`age45-54` + influenza*`age
45-54` + Northeast*`age45-54` + South*`age45-54` + West*`age45-54`)
               , plotit = "Observed", m=5)

## Analysis of Variance Table
##
## Response: num_deaths
##                                     Df Sum Sq Mean Sq F value Pr(>F)
## `age45-54`                         1 55638  55638  12.25 0.00056 ***
## adultrespdistSyn                   1  308   308   0.07 0.79473
## allothcondsandcauses              1 124916 124916  27.51 3.6e-07 ***
## alzheimer                          1 17801  17801  3.92 0.04892 *
## cardiacArrest                      1  9110   9110  2.01 0.15803
## cardiacarrhythmia                 1  1509   1509  0.33 0.56485
## cerebrovascula                     1   512    512  0.11 0.73733
## chroniclr                          1  1137   1137  0.25 0.61728
## covid19                            1 355946 355946  78.38 2.4e-16 ***
## diabetes                           1  3533   3533  0.78 0.37871
## heartfailure                       1   499    499  0.11 0.74051

```

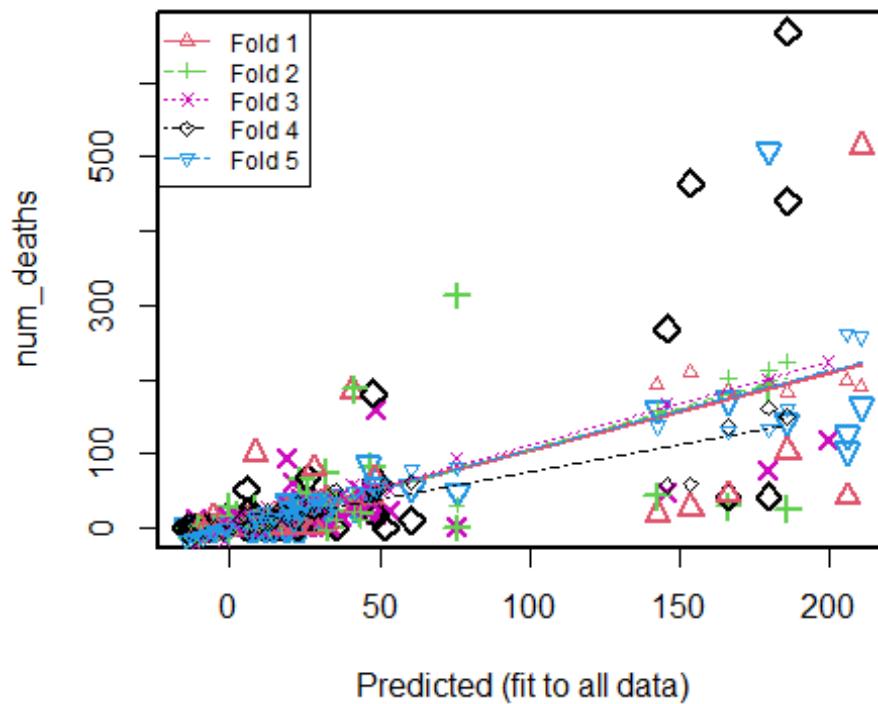
```

## hypertensive          1    196    196  0.04 0.83549
## influenza            1   7406   7406  1.63 0.20289
## injuryPoisoning     1    992    992  0.22 0.64066
## ischemicheart       1    500    500  0.11 0.74044
## malignant           1   1759   1759  0.39 0.53436
## obesity              1      5      5  0.00 0.97230
## othercirculatorysystem 1     46     46  0.01 0.91995
## otherrespsystem      1    203    203  0.04 0.83267
## renalfailure         1   4668   4668  1.03 0.31173
## conditionrespiratoryarrest 1   1303   1303  0.29 0.59269
## sepsis               1    176    176  0.04 0.84430
## vascular             1  12311  12311  2.71 0.10105
## Northeast            1    113    113  0.02 0.87462
## South                1   3229   3229  0.71 0.39996
## West                 1   2523   2523  0.56 0.45686
## `age45-54` :allothcondsandcauses 1   82454  82454 18.16 3.0e-05 ***
## `age45-54` :covid19        1   5193   5193  1.14 0.28603
## `age45-54` :influenza      1   383    383  0.08 0.77189
## `age45-54` :Northeast     1   1803   1803  0.40 0.52932
## `age45-54` :South         1    23     23  0.01 0.94314
## `age45-54` :West          1   1306   1306  0.29 0.59233
## Residuals            229 1039989 4541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in cv.lm(data = test, form.lm = formula(num_deaths ~ `age45-54` +
## :
## 
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

## Small symbols show cross-validation predicted values



```
##  
## fold 1  
## Observations in test set: 52  
##  
## Predicted      3.23 19.0 12.69 -3.47 40.9 47.869 154 11.5 29.87 19.26 3  
2.3  
## cvpred        8.67 19.5 6.14  1.95 27.0 38.099 209 16.4 34.88 24.38 4  
2.6  
## num_deaths    0.00  0.0  0.00  0.00 185.0 39.000   26  0.0 30.00 21.00  
0.0  
## CV residual -8.67 -19.5 -6.14 -1.95 158.0  0.901 -183 -16.4 -4.88 -3.38 -4  
2.6  
##  
## Predicted      5.65 28.9 -8.23 -5.17  8.39 -5.48 -10.1 186.2 28.9  9.77  
206  
## cvpred        8.74 39.7 -1.16 -3.54  8.56 -6.04 -11.4 181.7 39.7 14.79  
199  
## num_deaths    0.00 22.0  0.00  0.00 100.00 13.00   0.0 103.0 10.0  0.00  
41  
## CV residual -8.74 -17.7  1.16  3.54 91.44 19.04 11.4 -78.7 -29.7 -14.79  
-158  
##  
## Predicted      29.9 20.4 22.718  5.53 27.5  3.65 143 28.15 28.1 23.9  4  
2.3
```

```

## cvpred      34.9  11.8 14.838  7.63  25.7 -2.28  194 26.11 26.1 26.0  4
1.1
## num_deaths 24.0   0.0 14.000 11.00  11.0   0.00   18 17.00 14.0   0.0  3
1.0
## CV residual -10.9 -11.8 -0.838  3.37 -14.7   2.28 -176 -9.11 -12.1 -26.0 -1
0.1
##          1500 1517 1564 1586  1589 1600 1689 1892 1909 1914  1
965
## Predicted  3.44 -5.41 23.723 48.1  -9.72  211  31.9 -4.07  3.65 -9.99 31
.88
## cvpred     -1.43 -4.29 13.387 37.2 -12.88  190  42.4  3.83 -2.28 -2.77 42
.42
## num_deaths  0.00  0.00 14.000 64.0   0.00  514  10.0  0.00  0.00  0.00 39
.00
## CV residual 1.43  4.29  0.613 26.8  12.88  324 -32.4 -3.83  2.28  2.77 -3
.42
##          1994 2071 2138 2174 2236 2275 2304 2313
## Predicted  28.2  6.02 16.29 15.95 166  28.2 21.89 17.1
## cvpred     25.5 11.54  6.77 10.24 184  37.7  9.26 23.3
## num_deaths 79.0  0.00  0.00 15.00 44   0.0  0.00  0.0
## CV residual 53.5 -11.54 -6.77  4.76 -140 -37.7 -9.26 -23.3
##
## Sum of squares = 266979      Mean square = 5134      n = 52
##
## fold 2
## Observations in test set: 53
##          39   125   156   178   198   266   310   352   355   371
## Predicted 20.7 -3.475 12.877 179.7 75.6 35.1 14.5  9.77  5.78  6.02
## cvpred    25.6 -0.169  0.206 213.6 31.2 11.7 20.2  7.20 14.15 10.90
## num_deaths 0.0   0.000 11.000 190.0 313.0 22.0  0.0   0.00  0.00  0.00
## CV residual -25.6  0.169 10.794 -23.6 281.8 10.3 -20.2 -7.20 -14.15 -10.90
##          397   412   457   458   492   503   517   571   586   635
656
## Predicted -3.475 47.2  34.0 47.22 41.5  22.8 -5.17 -11.47 166 32.16  -0
.553
## cvpred     -0.169 43.2  35.8 43.15 22.5  25.2 -6.93 -9.51 203 -2.82 -10
.760
## num_deaths 0.000 82.0  15.0 53.00 42.0   0.0   0.00  0.00  29 74.00  28
.000
## CV residual 0.169 38.8 -20.8  9.85 19.5 -25.2  6.93  9.51 -174 76.82 38
.760
##          736   876   894   900   977   982  1003  1028 1034 1039  10
41
## Predicted -0.79 24.0  17.8 16.3 -2.04  33.0 32.16  8.46 23.6 -14.7 16.
97
## cvpred     -5.20 31.3  24.4 22.2 -6.60  43.1 -2.82  8.19 12.3 -13.6  1.
73
## num_deaths 0.00 10.0  11.0  0.0   0.00  0.0 46.00 25.00 65.0   0.0 12.
00
## CV residual 5.20 -21.3 -13.4 -22.2  6.60 -43.1 48.82 16.81 52.7  13.6 10.

```

```

27
##          1084 1204 1222 1239 1288 1321 1511 1559 1567 1645 1
675
## Predicted -8.23 75.6 5.20 29.9 -0.79 20.6 -0.629 26.0 8.78 143 -0.
215
## cvpred -12.65 31.2 -1.68 32.2 -5.20 22.8 0.483 25.5 9.70 158 3.
215
## num_deaths 0.00 0.0 0.00 14.0 0.00 14.0 0.000 47.0 12.00 43 0.
933
## CV residual 12.65 -31.2 1.68 -18.2 5.20 -8.8 -0.483 21.5 2.30 -115 -3.
933
##          1690 1734 1773 1873 1983 2008 2035 2061 2097 2211
## Predicted 186 12.7 -9.72 31.88 2.44 41.5 -5.17 43.41 -3.475 -10.1
## cvpred 225 10.7 -8.49 36.24 2.14 22.5 -6.93 10.62 -0.169 -15.8
## num_deaths 25 0.0 0.00 42.00 17.00 187.0 0.00 20.00 0.000 0.0
## CV residual -200 -10.7 8.49 5.76 14.86 164.5 6.93 9.38 0.169 15.8
##
## Sum of squares = 214150      Mean square = 4041      n = 53
##
## fold 3
## Observations in test set: 53
##          16 190 209 211 274 307 325 428 454 487 4
91
## Predicted 75.6 12.6 -11.4 18.99 53.4 -9.72 29.9 19.2 48.7 -1.71 28
.2
## cvpred 94.1 26.6 -23.1 8.87 50.2 -6.89 31.4 21.7 38.3 -1.74 33
.6
## num_deaths 0.0 14.0 13.0 93.00 23.0 0.00 0.0 17.0 158.0 0.00 12
.0
## CV residual -94.1 -12.6 36.1 84.13 -27.2 6.89 -31.4 -4.7 119.7 1.74 -21
.6
##          527 597 677 774 793 868 904 907 909 967 9
94
## Predicted 18.99 -5.41 20.62 48.7 5.96 28.5 46.2 23.9 146 -1.99 14
.5
## cvpred 8.87 -9.25 8.99 38.3 11.47 43.2 55.7 20.8 167 1.32 20
.9
## num_deaths 0.00 0.00 61.00 18.0 0.00 10.0 21.0 40.0 48 0.00 0
.0
## CV residual -8.87 9.25 52.01 -20.3 -11.47 -33.2 -34.7 19.2 -119 -1.32 -20
.9
##          1009 1014 1071 1213 1225 1327 1334 1375 1391 1438
1472
## Predicted 18.6 -0.498 3.115 -2.1 17.1 -0.749 -0.79 26.0 -2.04 24.0
12.6
## cvpred 14.4 -3.171 -0.176 -16.6 15.8 4.133 -1.13 34.1 6.32 33.6
26.6
## num_deaths 32.0 15.000 0.000 0.0 0.0 0.000 0.00 0.0 0.00 13.0
0.0
## CV residual 17.6 18.171 0.176 16.6 -15.8 -4.133 1.13 -34.1 -6.32 -20.6

```

```

-26.6
##          1523 1540 1554 1760 1805 1818 1836 1838 1936 1960 20
41
## Predicted -8.69 37.0 200 24.0 -2.04 32.9 47.2 43.11 30 9.77 -2
.1
## cvpred    -12.38 44.8 223 33.6 6.32 41.5 49.5 41.66 30 9.05 -16
.6
## num_deaths 0.00 15.0 119 11.0 0.00 0.0 33.0 49.00 13 0.00 0
.0
## CV residual 12.38 -29.8 -104 -22.6 -6.32 -41.5 -16.5 7.34 -17 -9.05 16
.6
##          2050 2052 2125 2152 2190 2225 2248 2252 2255
## Predicted 5.2 3.44 6.37 40.9 180 3.65 19.3 15.7 1.13
## cvpred    16.4 16.95 11.67 57.4 200 2.42 16.7 15.6 3.96
## num_deaths 0.0 0.00 0.00 25.0 78 13.00 0.0 0.0 0.00
## CV residual -16.4 -16.95 -11.67 -32.4 -122 10.58 -16.7 -15.6 -3.96
##
## Sum of squares = 90729      Mean square = 1712      n = 53
##
## fold 4
## Observations in test set: 52
##          7   94  126   159   183   217   313   320   382   385
434
## Predicted 18.7 47.2 9.77 3.835 -0.629 -14.6 -2.23 48.7 19.2 15.0 2
2.28
## cvpred    18.9 44.5 1.86 10.284 -4.323 -10.4 -4.71 45.2 12.2 10.4
8.87
## num_deaths 0.0 23.0 0.00 10.000 15.000 0.0 0.00 63.0 0.0 0.0 2
5.00
## CV residual -18.9 -21.5 -1.86 -0.284 19.323 10.4 4.71 17.8 -12.2 -10.4 1
6.13
##          550   628   630   647  680    761   808   881   936   944   96
2
## Predicted 23.36 35.1 12.7 5.53 186 -12.97 11.53 60.7 28.21 28.1 52.
0
## cvpred    17.87 50.7 21.7 1.61 150 -6.68 3.86 61.6 22.56 20.2 58.
7
## num_deaths 21.00 0.0 0.0 52.00 440 0.00 0.00 10.0 17.00 33.0 0.
0
## CV residual 3.13 -50.7 -21.7 50.39 290 6.68 -3.86 -51.6 -5.56 12.8 -58.
7
##          963   965   992   1083  1136   1138   1171   1197   1221   130
8
## Predicted 21.9 -2.28 16.3 -12.97 -0.74 166.3 30.1 -5.1181 -12.85 -6.7
3
## cvpred    17.7 -6.52 18.0 -6.68 -4.51 135.6 22.8 -0.0488 -8.42 -8.6
8
## num_deaths 0.0 0.00 0.0 0.00 0.00 42.0 12.0 0.0000 0.00 0.0
## CV residual -17.7 6.52 -18.0 6.68 4.51 -93.6 -10.8 0.0488 8.42 8.6

```

```

8
##          1315 1421 1461 1490 1506 1513 1544 1562 1626 1636 16
48
## Predicted -14.73   26 145.8 16.0 180 22.8 5.2 47.9 8.46 -8.23 48
.7
## cvpred     -8.69   31 57.3 21.8 162 32.2 16.8 45.3 9.24 -9.46 45
.2
## num_deaths  0.00   67 268.0 10.0 42 15.0 0.0 180.0 16.00 0.00 17
.0
## CV residual 8.69   36 210.7 -11.8 -120 -17.2 -16.8 134.7 6.76 9.46 -28
.2
##          1775 1776 1814 1860 1883 1979 2012 2013 2019
## Predicted -11.5 14.5 17.8 13.03 18.6 6.0176 186 153.8 22.8
## cvpred     -11.7 16.0 18.8 4.64 20.6 0.0981 149 58.7 32.2
## num_deaths  0.0 0.0 0.0 0.00 0.0 22.0000 668 464.0 0.0
## CV residual 11.7 -16.0 -18.8 -4.64 -20.6 21.9019 519 405.3 -32.2
##
## Sum of squares = 624385      Mean square = 12007      n = 52
##
## fold 5
## Observations in test set: 52
##          5    12    26    31    123    152    176    179    204    224    3
06
## Predicted 22.8 23.7 6.70 -12.8 -1.711 75.6 12.69 142.7 21.29 180 32
.9
## cvpred     21.5 18.2 -4.55 -13.4 0.468 81.6 3.21 137.7 6.87 134 35
.7
## num_deaths 0.0 0.0 0.00 0.0 0.000 48.0 15.00 160.0 34.00 509 10
.0
## CV residual -21.5 -18.2 4.55 13.4 -0.468 -33.6 11.79 22.3 27.13 375 -25
.7
##          411    486    533    537    757    840    849    858    871    910
946
## Predicted 33.97 28.1 -14.6 -5.36 18.9 19.2 30.1 41.5 21.87 210.7 1
1.5
## cvpred     35.41 38.8 -17.8 -5.78 27.7 27.1 40.8 49.6 24.09 257.4 2
1.3
## num_deaths 40.00 25.0 0.0 0.00 0.0 17.0 18.0 28.0 22.00 164.0
0.0
## CV residual 4.59 -13.8 17.8 5.78 -27.7 -10.1 -22.8 -21.6 -2.09 -93.4 -2
1.3
##          949    954    985   1046   1090   1091   1207   1216   1287   1293
## Predicted 28.2 186.0 -8.22 166.3 -0.740 20.623 -5.17 6.70 -5.53 -8.81
## cvpred     28.8 160.1 -7.96 131.6 0.452 19.717 -7.57 -4.55 -4.32 -13.11
## num_deaths 14.0 143.0 0.00 172.0 0.000 20.000 0.00 0.00 0.00 0.00
## CV residual -14.8 -17.1 7.96 40.4 -0.452 0.283 7.57 4.55 4.32 13.11
##          1312   1362   1428   1433   1456   1772   1815   1866   1876   1879
1929
## Predicted 8.39 -9.99 24.02 60.7 46.2 32.90 22.1 11.5 206 -0.749
29.9

```

```

## cvpred      10.33 -11.58 21.95  78.2 46.4 35.66  32.5 21.3 261 -1.063
32.4
## num_deaths   0.00   0.00 29.00  53.0 85.0 27.00   0.0   0.0 104  0.000
19.0
## CV residual -10.33  11.58  7.05 -25.2 38.6 -8.66 -32.5 -21.3 -157  1.063 -
13.4
##           1957  2010  2062 2106  2117  2122  2140  2195 2289
## Predicted  -1.711 19.02 48.68 206   8.78  19.1  14.5 -0.749 18.7
## cvpred      0.468 28.93 53.48 261  14.81  26.4  13.8 -1.063 18.5
## num_deaths   0.000 33.00 55.00 126   0.00   0.0   0.0  0.000  0.0
## CV residual -0.468  4.07  1.52 -135 -14.81 -26.4 -13.8  1.063 -18.5
##
## Sum of squares = 206856     Mean square = 3978     n = 52
##
## Overall (Sum over all 52 folds)
## ms
## 5355

```

Final model with interaction terms

Prunned *age45-54:chroniclr*, *age45-54:obesity*

```

##, age45-54:heartfailure, age45-54:cerebrovascula ##, ##age45-
54:otherrespsystem,age45-54:malignant , age45-54:injuryPoisoning, age45-
54:cardiacarrhythmia, ##age45-54:conditionrespiratoryarrest, age45-54:vascular ##, age45-
54:alzheimer ,age45-54:renalfailure,age45-54:othercirculatorysystem

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothc
ondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` + covid19*`age45-54` + influenza*`age
45-54` + Northeast*`age45-54` + South*`age45-54` + West*`age45-54` , data =
covidcleansed)
summary(model.interactionH)

##
## Call:
## lm(formula = num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandca
uses +
##   alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
##   chroniclr + covid19 + diabetes + heartfailure + hypertensive +
##   influenza + injuryPoisoning + ischemicheart + malignant +
##   obesity + othercirculatorysystem + otherrespsystem + renalfailure +

```

```

## conditionrespiratoryarrest + sepsis + vascular + Northeast +
## South + West + allothcondsandcauses * `age45-54` + covid19 *
## `age45-54` + influenza * `age45-54` + Northeast * `age45-54` +
## South * `age45-54` + West * `age45-54`, data = covidcleansed)
##
## Residuals:
##      Min    1Q Median    3Q   Max
## -195.1 -20.3  -4.3   6.9 980.9
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 45.499    10.028   4.54  6.2e-06 ***
## `age45-54`                   0.231     8.411   0.03  0.97812  
## adultrespdistSyn            -22.422   11.974  -1.87  0.06134 .  
## allothcondsandcauses        -9.154    13.698  -0.67  0.50410  
## alzheimer                    -57.026   10.698  -5.33  1.1e-07 *** 
## cardiacArrest                -28.234   12.889  -2.19  0.02866 *  
## cardiacarrhythmia           -53.182   12.531  -4.24  2.3e-05 *** 
## cerebrovascula              -52.340   13.247  -3.95  8.2e-05 *** 
## chroniclr                     -49.287   12.588  -3.92  9.5e-05 *** 
## covid19                       29.636   13.811   2.15  0.03207 *  
## diabetes                      -20.360   12.092  -1.68  0.09247 .  
## heartfailure                  -52.240   13.251  -3.94  8.5e-05 *** 
## hypertensive                  -25.599   12.451  -2.06  0.03997 *  
## influenza                      -6.130    14.846  -0.41  0.67973  
## injuryPoisoning               -50.232   12.326  -4.08  4.9e-05 *** 
## ischemicheart                 -47.497   12.899  -3.68  0.00024 *** 
## malignant                     -52.355   13.087  -4.00  6.7e-05 *** 
## obesity                        -29.751   12.381  -2.40  0.01640 *  
## othercirculatorysystem        -45.236   13.041  -3.47  0.00054 *** 
## otherrespystem                -51.274   13.243  -3.87  0.00011 *** 
## renalfailure                  -39.397   12.332  -3.19  0.00143 **  
## conditionrespiratoryarrest   -53.571   12.921  -4.15  3.6e-05 *** 
## sepsis                         -33.596   12.318  -2.73  0.00647 ** 
## vascular                       -56.005   11.032  -5.08  4.4e-07 *** 
## Northeast                      4.799    8.520   0.56  0.57337  
## South                          8.604    7.564   1.14  0.25556  
## West                           3.640    7.600   0.48  0.63203  
## `age45-54`:allothcondsandcauses 59.403   15.780  3.76  0.00017 *** 
## `age45-54`:covid19             103.492   15.705  6.59  6.3e-11 *** 
## `age45-54`:influenza          50.918    17.165  2.97  0.00307 ** 
## `age45-54`:Northeast           21.467    11.802  1.82  0.06916 .  
## `age45-54`:South               16.661    10.479  1.59  0.11208  
## `age45-54`:West                18.362    10.981  1.67  0.09474 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.9 on 1331 degrees of freedom
## Multiple R-squared:  0.283, Adjusted R-squared:  0.266 
## F-statistic: 16.4 on 32 and 1331 DF,  p-value: <2e-16

```

## Residual Analysis

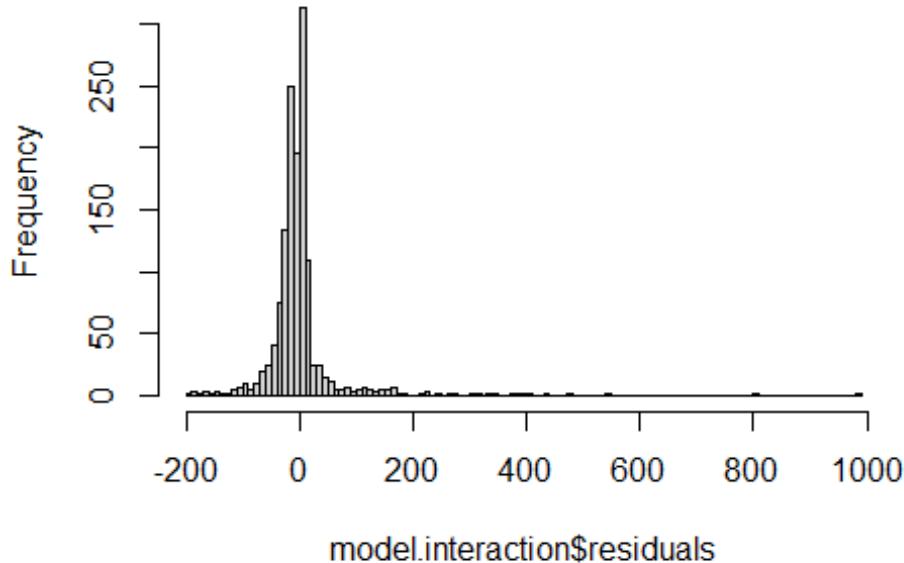
Verying the assumption that the sum of the residuals should add up to zero.

```
model.interactionH <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothc  
ondsandcauses +  
    alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +  
    chroniclr + covid19 + diabetes + heartfailure + hypertensive +  
    influenza + injuryPoisoning + ischemicheart + malignant +  
    obesity + othercirculatorysystem + otherrespsystem + renalfailure +  
    conditionrespiratoryarrest + sepsis + vascular + Northeast +  
    South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` + i  
nfluenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` + West*`age45  
-54`, data = covidcleansed)  
sum(model.interactionH$residuals)  
  
## [1] -1.81e-14
```

##Histogram of residuals

```
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2  
model.interaction <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothc  
ondsandcauses +  
    alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +  
    chroniclr + covid19 + diabetes + heartfailure + hypertensive +  
    influenza + injuryPoisoning + ischemicheart + malignant +  
    obesity + othercirculatorysystem + otherrespsystem + renalfailure +  
    conditionrespiratoryarrest + sepsis + vascular + Northeast +  
    South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` + i  
nfluenza*`age45-54`  
    + Northeast*`age45-54` + South*`age45-54` + West*`age45-54`, data = covid  
cleansed)  
hist(model.interaction$residuals, breaks = 100)
```

## Histogram of model.interaction\$residuals

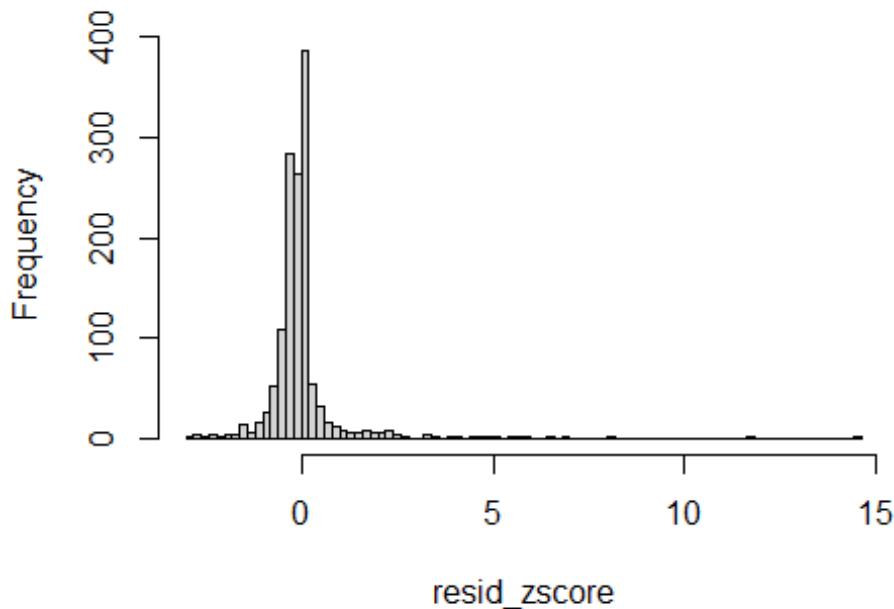


```
##idea about the distribution in the curve
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interaction <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West +
  ##
  allothcondsandcauses*`age45-54` + covid19*`age45-54` + influenza*`age
  45-54` + Northeast*`age45-54` + South*`age45-54` + West*`age45-54`, data =
  covidcleansed)

mean = mean(model.interaction$residuals) #

## This is the standard deviation of the residuals which I am to save as sd.
sd = sd(model.interaction$residuals)
resid_zscore = (model.interaction$residuals - mean)/sd
hist(resid_zscore, breaks = 100)
```

## Histogram of resid\_zscore



```
##Verifying the independence of residuals ## using Durbin Watson test on the model.
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interaction <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` + i
nfluenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` + West*`age45
-54` , data = covidcleansed)
library(car)

## Warning: package 'car' was built under R version 4.0.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.0.3

##
## Attaching package: 'car'

## The following object is masked from 'package:DAAG':
##
##      vif
```

```

## The following object is masked from 'package:dplyr':
##
##      recode
## The following object is masked from 'package:purrr':
##
##      some
durbinWatsonTest(model.interaction)

##   lag Autocorrelation D-W Statistic p-value
##     1           0.42          1.16       0
## Alternative hypothesis: rho != 0

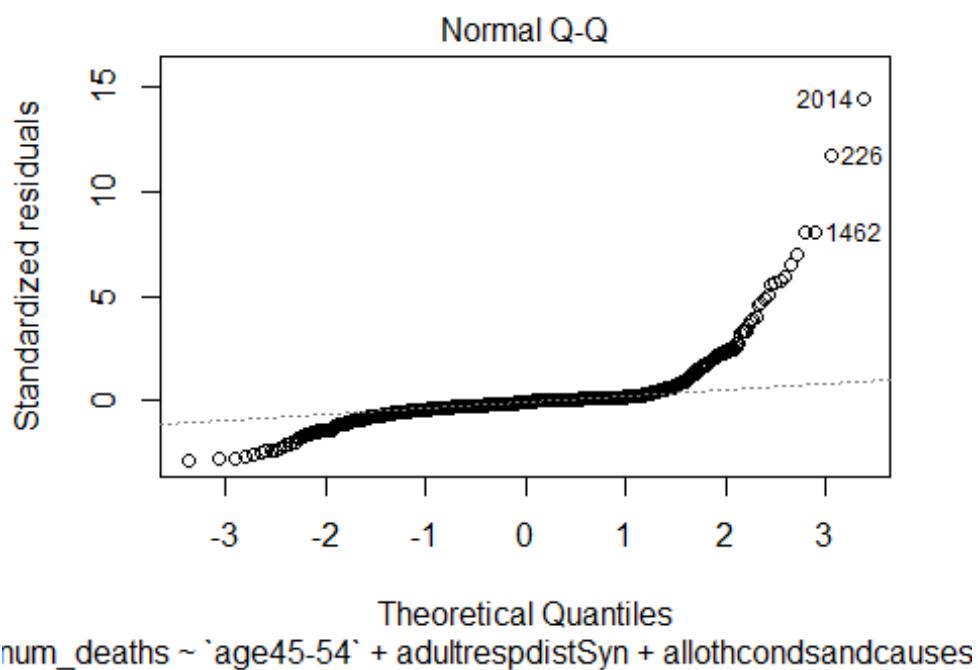
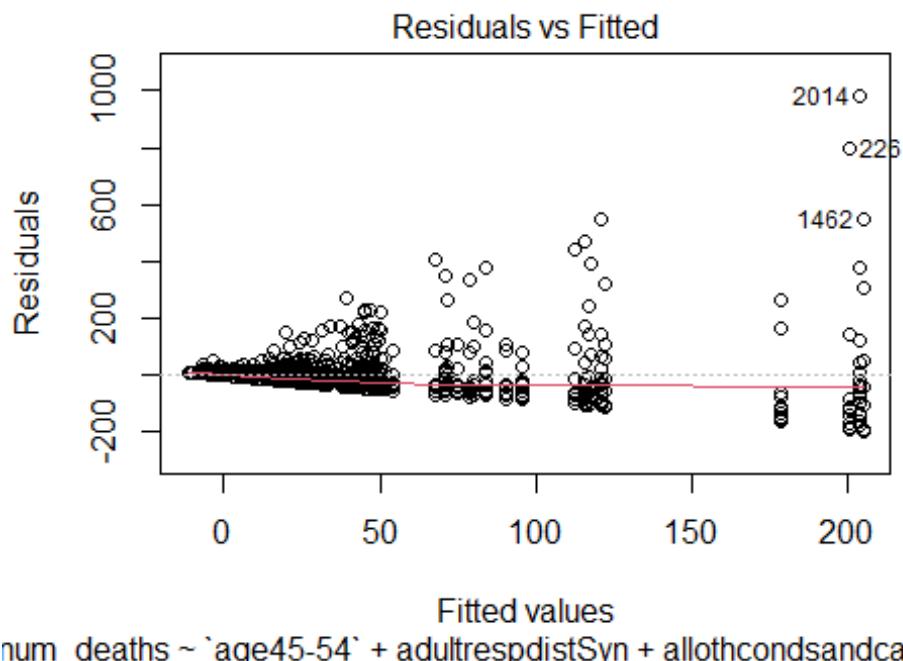
```

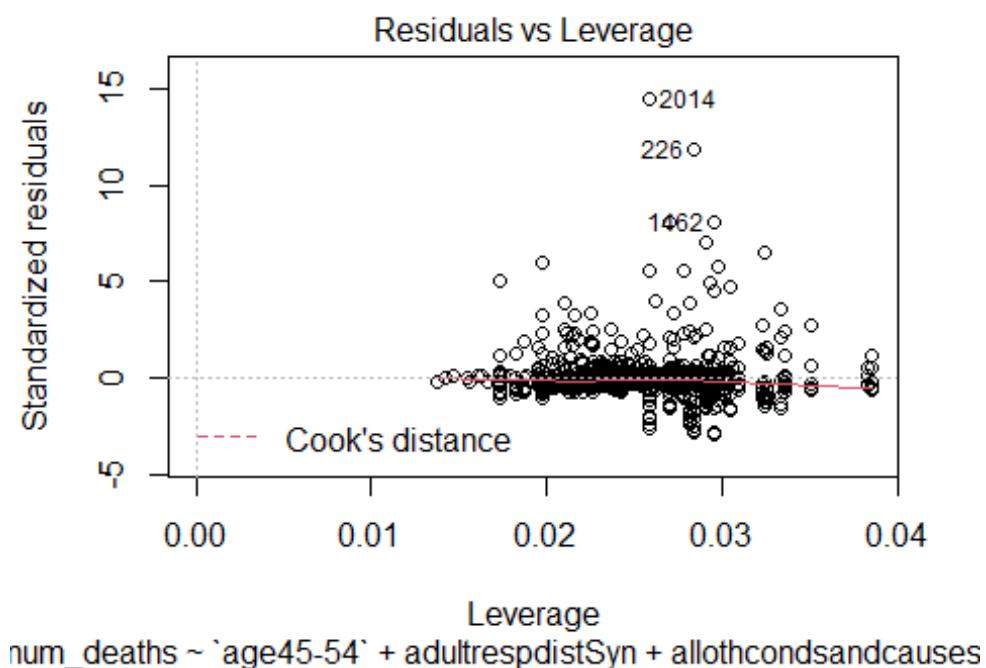
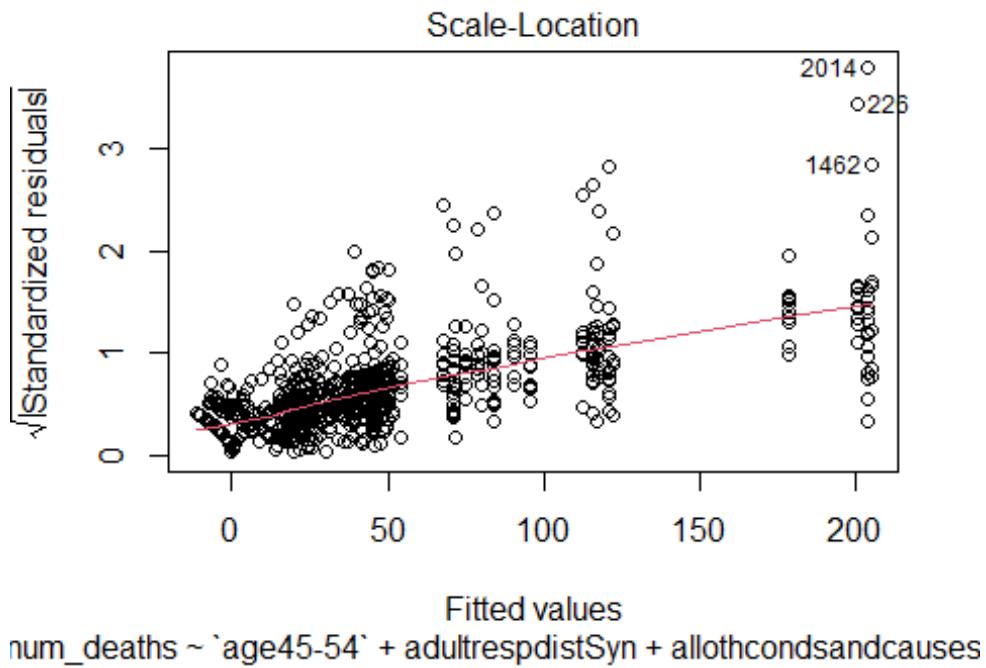
### Residual plots

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(num_deaths ~ `age45-54` + adultrespdistSyn + allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` + i
nfluenza*`age45-54` + Northeast*`age45-54` + South*`age45-54` + West*`age45
-54` , data = covidcleansed)
prediction <- predict(model.interaction, covidcleansed)
plot(model.interactionH)

```





Removing specific rows in r

```
## Adding an index column to my data set
covidcleansed$generated_uid <- 1:nrow(covidcleansed)
covidcleansed <- covidcleansed[-c(1203, 880, 153),]
```

## Stabilized model

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
+ allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses * `age45-54` + covid19 * `age45-54` + i
nfluenza * `age45-54` + Northeast * `age45-54` + South * `age45-54` + West * `age45
-54`, data = covidcleansed)
prediction <- predict(model.interactionH, covidcleansed)
summary(model.interactionH)

##
## Call:
## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + South + West + allothcondsandcauses * `age45-54` +
##     covid19 * `age45-54` + influenza * `age45-54` + Northeast *
##     `age45-54` + South * `age45-54` + West * `age45-54`, data = covidclean
sed)
##
## Residuals:
##   Min     1Q Median     3Q    Max
## -3.926 -0.808 -0.048  0.722  3.349
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.4972    0.1837  13.60 < 2e-16 ***
## `age45-54`                  0.8701    0.1541   5.65 2.0e-08 ***
## adultrespdistSyn            -0.3737    0.2193  -1.70  0.08858 .
## allothcondsandcauses        0.3188    0.2509   1.27  0.20410
## alzheimer                  -3.1180    0.1959 -15.91 < 2e-16 ***
## cardiacArrest               -0.9253    0.2361  -3.92 9.3e-05 ***
## cardiacarrhythmia           -2.4215    0.2295 -10.55 < 2e-16 ***
## cerebrovascula              -2.3481    0.2426  -9.68 < 2e-16 ***
## chroniclr                   -2.0561    0.2306  -8.92 < 2e-16 ***
## covid19                      1.0067    0.2530   3.98 7.3e-05 ***
## diabetes                     -0.5656    0.2215  -2.55  0.01076 *
## heartfailure                 -2.2268    0.2427  -9.18 < 2e-16 ***
## hypertensive                 -0.8352    0.2280  -3.66  0.00026 ***
## influenza                     0.3091    0.2719   1.14  0.25585
## injuryPoisoning              -2.0579    0.2258  -9.12 < 2e-16 ***
## ischemicheart                -2.0304    0.2363  -8.59 < 2e-16 ***

```

```

## malignant           -2.3444    0.2397   -9.78 < 2e-16 ***
## obesity            -0.8527    0.2268   -3.76  0.00018 ***
## othercirculatorysystem -1.4859    0.2388   -6.22  6.6e-10 ***
## otherrespsystem    -2.1720    0.2425   -8.96 < 2e-16 ***
## renalfailure       -1.2481    0.2259   -5.53  3.9e-08 ***
## conditionrespiratoryarrest -2.5912    0.2366   -10.95 < 2e-16 ***
## sepsis              -0.9008    0.2256   -3.99  6.9e-05 ***
## vascular            -3.0623    0.2021   -15.16 < 2e-16 ***
## Northeast           0.0898    0.1561    0.58  0.56495
## South               0.5445    0.1385    3.93  8.9e-05 ***
## West                -0.1008    0.1392   -0.72  0.46925
## `age45-54` :allothcondsandcauses 0.3091    0.2890    1.07  0.28499
## `age45-54` :covid19      -0.1648    0.2925   -0.56  0.57331
## `age45-54` :influenza     0.1692    0.3144    0.54  0.59063
## `age45-54` :Northeast     0.1609    0.2164    0.74  0.45716
## `age45-54` :South        0.0141    0.1920    0.07  0.94138
## `age45-54` :West         0.0588    0.2013    0.29  0.77032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1328 degrees of freedom
## Multiple R-squared:  0.561, Adjusted R-squared:  0.55
## F-statistic:  53 on 32 and 1328 DF,  p-value: <2e-16

##plot(model.interactionH)

```

Stabilized model

Prune *age45-54*:South

Adjusted R-squared: 0.55

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn
+ allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` + i
nfluenza*`age45-54` + Northeast*`age45-54` + West*`age45-54`, data = covidc
leansed)
summary(model.interactionH)

##
## Call:
## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular + Northeast +
##     South + West + allothcondsandcauses*`age45-54` + covid19*`age45-54` + i
nfluenza*`age45-54` + Northeast*`age45-54` + West*`age45-54`)

```

```

##      renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##      Northeast + South + West + allothcondsandcauses * `age45-54` +
##      covid19 * `age45-54` + influenza * `age45-54` + Northeast *
##      `age45-54` + West * `age45-54`, data = covidcleansed)
##
## Residuals:
##      Min    1Q Median    3Q   Max
## -3.923 -0.805 -0.049  0.722  3.349
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                2.4927    0.1729   14.42 < 2e-16 ***
## `age45-54`                  0.8786    0.1005    8.75 < 2e-16 ***
## adultrespdistSyn          -0.3737    0.2192   -1.70  0.08851 .  
## allothcondsandcauses       0.3187    0.2508    1.27  0.20399    
## alzheimer                 -3.1179    0.1959   -15.92 < 2e-16 ***
## cardiacArrest              -0.9246    0.2358   -3.92 9.3e-05 ***
## cardiacarrhythmia         -2.4210    0.2293   -10.56 < 2e-16 ***
## cerebrovascula            -2.3477    0.2425   -9.68 < 2e-16 ***
## chroniclr                  -2.0559    0.2304   -8.92 < 2e-16 ***
## covid19                     1.0069    0.2529    3.98 7.2e-05 ***
## diabetes                   -0.5656    0.2214   -2.55  0.01073 *  
## heartfailure               -2.2261    0.2424   -9.18 < 2e-16 ***
## hypertensive                -0.8348    0.2279   -3.66  0.00026 *** 
## influenza                   0.3085    0.2717    1.14  0.25634    
## injuryPoisoning            -2.0581    0.2257   -9.12 < 2e-16 ***
## ischemicheart              -2.0300    0.2361   -8.60 < 2e-16 ***
## malignant                  -2.3439    0.2395   -9.79 < 2e-16 ***
## obesity                     -0.8524    0.2266   -3.76  0.00018 *** 
## othercirculatorysystem     -1.4859    0.2388   -6.22 6.5e-10 ***
## otherrespystem              -2.1717    0.2424   -8.96 < 2e-16 ***
## renalfailure                -1.2479    0.2258   -5.53 3.9e-08 ***
## conditionrespiratoryarrest -2.5906    0.2364   -10.96 < 2e-16 ***
## sepsis                      -0.9007    0.2255   -3.99 6.9e-05 *** 
## vascular                    -3.0622    0.2020  -15.16 < 2e-16 ***
## Northeast                   0.0942    0.1444    0.65  0.51442    
## South                       0.5519    0.0962    5.74  1.2e-08 ***
## West                        -0.0964    0.1261   -0.76  0.44468    
## `age45-54`:allothcondsandcauses 0.3092    0.2889    1.07  0.28474    
## `age45-54`:covid19          -0.1650    0.2924   -0.56  0.57276    
## `age45-54`:influenza        0.1701    0.3140    0.54  0.58808    
## `age45-54`:Northeast        0.1523    0.1817    0.84  0.40205    
## `age45-54`:West             0.0502    0.1638    0.31  0.75935    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1329 degrees of freedom
## Multiple R-squared:  0.561, Adjusted R-squared:  0.55
## F-statistic: 54.7 on 31 and 1329 DF, p-value: <2e-16

```

Pruned age45-54:West

Adjusted R-squared: 0.551

```
covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
+ allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast +
  South + West + allothcondsandcauses * `age45-54` + covid19 * `age45-54` + influenza * `age45-54` + Northeast * `age45-54`, data = covidcleansed)
summary(model.interactionH)

##
## Call:
## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + South + West + allothcondsandcauses * `age45-54` +
##     covid19 * `age45-54` + influenza * `age45-54` + Northeast *
##     `age45-54`, data = covidcleansed)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -3.931 -0.810 -0.063  0.721  3.368
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.4835    0.1703   14.59 < 2e-16 ***
## `age45-54`                  0.8956    0.0838   10.69 < 2e-16 ***
## adultrespdistSyn          -0.3737    0.2191   -1.71  0.08834 .
## allothcondsandcauses       0.3208    0.2506    1.28  0.20082
## alzheimer                 -3.1167    0.1957  -15.92 < 2e-16 ***
## cardiacArrest              -0.9248    0.2357   -3.92 9.2e-05 ***
## cardiacarrhythmia          -2.4206    0.2293  -10.56 < 2e-16 ***
## cerebrovascula             -2.3463    0.2423   -9.68 < 2e-16 ***
## chroniclr                  -2.0548    0.2303   -8.92 < 2e-16 ***
## covid19                     1.0097    0.2526    4.00 6.8e-05 ***
## diabetes                   -0.5660    0.2213   -2.56  0.01065 *
## heartfailure               -2.2254    0.2423   -9.18 < 2e-16 ***
## hypertensive                -0.8348    0.2278   -3.66  0.00026 ***
## influenza                   0.3100    0.2716    1.14  0.25378
## injuryPoisoning            -2.0565    0.2255  -9.12 < 2e-16 ***
## ischemicheart              -2.0318    0.2359  -8.61 < 2e-16 ***
## malignant                  -2.3428    0.2394  -9.79 < 2e-16 ***
```

```

## obesity -0.8514 0.2265 -3.76 0.00018 ***
## othercirculatorysystem -1.4860 0.2387 -6.23 6.4e-10 ***
## otherrespsystem -2.1709 0.2423 -8.96 < 2e-16 ***
## renalfailure -1.2471 0.2257 -5.53 3.9e-08 ***
## conditionrespiratoryarrest -2.5889 0.2363 -10.96 < 2e-16 ***
## sepsis -0.9002 0.2254 -3.99 6.9e-05 ***
## vascular -3.0607 0.2018 -15.16 < 2e-16 ***
## Northeast 0.1025 0.1418 0.72 0.46995
## South 0.5514 0.0961 5.74 1.2e-08 ***
## West -0.0731 0.1005 -0.73 0.46725
## `age45-54` :allothcondsandcauses 0.3061 0.2886 1.06 0.28903
## `age45-54` :covid19 -0.1688 0.2920 -0.58 0.56326
## `age45-54` :influenza 0.1674 0.3138 0.53 0.59377
## `age45-54` :Northeast 0.1358 0.1735 0.78 0.43385
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1330 degrees of freedom
## Multiple R-squared: 0.561, Adjusted R-squared: 0.551
## F-statistic: 56.6 on 30 and 1330 DF, p-value: <2e-16

```

##Prunned age45-54:influenza ## Adjusted R-squared: 0.551

```

covidcleansed$`age45-54$Q` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
+ allothcondsandcauses +
alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
chroniclr + covid19 + diabetes + heartfailure + hypertensive +
influenza + injuryPoisoning + ischemicheart + malignant +
obesity + othercirculatorysystem + otherrespsystem + renalfailure +
conditionrespiratoryarrest + sepsis + vascular + Northeast +
South + West + allothcondsandcauses * `age45-54` + covid19 * `age45-54` + N
ortheast * `age45-54`, data = covidcleansed)
summary(model.interactionH)

##
## Call:
## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + South + West + allothcondsandcauses * `age45-54` +
##     covid19 * `age45-54` + Northeast * `age45-54`, data = covidcleansed)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -3.935 -0.805 -0.059  0.726  3.364
## 
```

```

## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.4791   0.1700   14.58 < 2e-16 ***
## `age45-54`                  0.9059   0.0815   11.11 < 2e-16 ***
## adultrespdistSyn          -0.3737   0.2191   -1.71  0.08830 .
## allothcondsandcauses       0.3260   0.2504    1.30  0.19309
## alzheimer                 -3.1163   0.1957  -15.92 < 2e-16 ***
## cardiacArrest              -0.9250   0.2356   -3.93 9.1e-05 ***
## cardiacarrhythmia          -2.4196   0.2292  -10.56 < 2e-16 ***
## cerebrovascula             -2.3450   0.2423   -9.68 < 2e-16 ***
## chroniclr                  -2.0544   0.2303   -8.92 < 2e-16 ***
## covid19                     1.0149   0.2523    4.02 6.1e-05 ***
## diabetes                    -0.5670   0.2212   -2.56  0.01049 *
## heartfailure               -2.2239   0.2422   -9.18 < 2e-16 ***
## hypertensive                -0.8350   0.2277   -3.67  0.00026 ***
## influenza                   0.3998   0.2130    1.88  0.06077 .
## injuryPoisoning            -2.0562   0.2255   -9.12 < 2e-16 ***
## ischemicheart               -2.0307   0.2359   -8.61 < 2e-16 ***
## malignant                  -2.3420   0.2393   -9.79 < 2e-16 ***
## obesity                     -0.8517   0.2265   -3.76  0.00018 ***
## othercirculatorysystem      -1.4856   0.2386   -6.23 6.4e-10 ***
## otherrespsystem              2.1701   0.2423   -8.96 < 2e-16 ***
## renalfailure                -1.2483   0.2256   -5.53 3.8e-08 ***
## conditionrespiratoryarrest -2.5885   0.2362  -10.96 < 2e-16 ***
## sepsis                      -0.9002   0.2254   -3.99 6.8e-05 ***
## vascular                    -3.0598   0.2018  -15.16 < 2e-16 ***
## Northeast                   0.1025   0.1418    0.72  0.46991
## South                       0.5501   0.0961    5.73 1.3e-08 ***
## West                        -0.0743   0.1005   -0.74  0.45989
## `age45-54`:allothcondsandcauses 0.2962   0.2880    1.03  0.30377
## `age45-54`:covid19          -0.1788   0.2914   -0.61  0.53963
## `age45-54`:Northeast        0.1335   0.1734    0.77  0.44127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1331 degrees of freedom
## Multiple R-squared:  0.561, Adjusted R-squared:  0.551
## F-statistic: 58.5 on 29 and 1331 DF, p-value: <2e-16

```

##Pruned age45-54:covid19 ## Adjusted R-squared: 0.551

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
+ allothcondsandcauses +
+ alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
+ chroniclr + covid19 + diabetes + heartfailure + hypertensive +
+ influenza + injuryPoisoning + ischemicheart + malignant +
+ obesity + othercirculatorysystem + otherrespsystem + renalfailure +
+ conditionrespiratoryarrest + sepsis + vascular + Northeast +
+ South + West + allothcondsandcauses*`age45-54` + Northeast*`age45-54` , da

```

```

ta = covidcleansed)
summary(model.interactionH)

##
## Call:
## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + South + West + allothcondsandcauses * `age45-54` +
##     Northeast * `age45-54`, data = covidcleansed)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -3.930 -0.810 -0.064  0.720  3.369
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.4852    0.1697   14.65 < 2e-16 ***
## `age45-54`                  0.8938    0.0791   11.30 < 2e-16 ***
## adultrespdistSyn          -0.3737    0.2190   -1.71  0.08819 .
## allothcondsandcauses       0.3199    0.2501    1.28  0.20115
## alzheimer                 -3.1169    0.1956  -15.93 < 2e-16 ***
## cardiacArrest              -0.9247    0.2356   -3.93 9.1e-05 ***
## cardiacarrhythmia          -2.4208    0.2291  -10.57 < 2e-16 ***
## cerebrovascula             -2.3465    0.2422   -9.69 < 2e-16 ***
## chroniclr                  -2.0549    0.2302   -8.93 < 2e-16 ***
## covid19                     0.9252    0.2055    4.50 7.3e-06 ***
## diabetes                   -0.5659    0.2212   -2.56 0.01062 *
## heartfailure               -2.2255    0.2422   -9.19 < 2e-16 ***
## hypertensive                -0.8347    0.2277   -3.67 0.00026 ***
## influenza                   0.3998    0.2130    1.88 0.06073 .
## injuryPoisoning            -2.0566    0.2254   -9.12 < 2e-16 ***
## ischemicheart              -2.0321    0.2358   -8.62 < 2e-16 ***
## malignant                  -2.3430    0.2393   -9.79 < 2e-16 ***
## obesity                     -0.8514    0.2264   -3.76 0.00018 ***
## othercirculatorysystem     -1.4861    0.2386   -6.23 6.3e-10 ***
## otherrespsystem             -2.1711    0.2422   -8.96 < 2e-16 ***
## renalfailure               -1.2468    0.2255   -5.53 3.9e-08 ***
## conditionrespiratoryarrest -2.5891    0.2362  -10.96 < 2e-16 ***
## sepsis                      -0.9002    0.2253   -4.00 6.8e-05 ***
## vascular                   -3.0609    0.2017  -15.17 < 2e-16 ***
## Northeast                  0.1014    0.1417    0.72 0.47418
## South                       0.5508    0.0960    5.73 1.2e-08 ***
## West                        -0.0743    0.1005   -0.74 0.45946
## `age45-54`:allothcondsandcauses 0.3077    0.2873    1.07 0.28428
## `age45-54`:Northeast        0.1368    0.1732    0.79 0.42994
## ---

```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1332 degrees of freedom
## Multiple R-squared:  0.56,   Adjusted R-squared:  0.551
## F-statistic: 60.6 on 28 and 1332 DF,  p-value: <2e-16

```

Pruned *age45-54:Northeast*

Adjusted R-squared: 0.555

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
+ allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast + West +
  South + allothcondsandcauses * `age45-54`, data = covidcleansed)
summary(model.interactionH)

##
## Call:
## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + West + South + allothcondsandcauses * `age45-54`,
##     data = covidcleansed)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -3.941 -0.806 -0.055  0.726  3.357
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                2.4699    0.1685  14.65 < 2e-16 ***
## `age45-54`                  0.9209    0.0712  12.93 < 2e-16 ***
## adultrespdistSyn            -0.3739    0.2190   -1.71  0.08800 .
## allothcondsandcauses        0.3208    0.2501   1.28  0.19977  
## alzheimer                   -3.1158    0.1956  -15.93 < 2e-16 ***
## cardiacArrest                -0.9259    0.2356   -3.93 8.9e-05 ***
## cardiacarrhythmia            -2.4158    0.2290  -10.55 < 2e-16 ***
## cerebrovascula               -2.3420    0.2421   -9.67 < 2e-16 ***
## chroniclr                     -2.0516    0.2302  -8.91 < 2e-16 ***
## covid19                        0.9261    0.2055   4.51 7.1e-06 ***
## diabetes                      -0.5666    0.2212  -2.56  0.01052 *  
## heartfailure                  -2.2250    0.2421  -9.19 < 2e-16 ***
## hypertensive                  -0.8324    0.2276  -3.66  0.00027 ***

```

```

## influenza 0.4007 0.2130 1.88 0.06009 .
## injuryPoisoning -2.0536 0.2253 -9.11 < 2e-16 ***
## ischemicheart -2.0277 0.2357 -8.60 < 2e-16 ***
## malignant -2.3356 0.2390 -9.77 < 2e-16 ***
## obesity -0.8500 0.2264 -3.75 0.00018 ***
## othercirculatorysystem -1.4840 0.2385 -6.22 6.6e-10 ***
## otherrespsystem -2.1663 0.2421 -8.95 < 2e-16 ***
## renalfailure -1.2470 0.2255 -5.53 3.9e-08 ***
## conditionrespiratoryarrest -2.5866 0.2361 -10.96 < 2e-16 ***
## sepsis -0.9018 0.2253 -4.00 6.6e-05 ***
## vascular -3.0608 0.2017 -15.18 < 2e-16 ***
## Northeast 0.1741 0.1078 1.62 0.10647
## West -0.0728 0.1004 -0.72 0.46866
## South 0.5502 0.0960 5.73 1.2e-08 ***
## `age45-54`:allothcondsandcauses 0.3058 0.2872 1.06 0.28721
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1333 degrees of freedom
## Multiple R-squared: 0.56, Adjusted R-squared: 0.551
## F-statistic: 62.9 on 27 and 1333 DF, p-value: <2e-16

```

Prunned *age45-54*:allothcondsandcauses

Adjusted R-squared: 0.555

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
+ allothcondsandcauses +
+ alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
+ chroniclr + covid19 + diabetes + heartfailure + hypertensive +
+ influenza + injuryPoisoning + ischemicheart + malignant +
+ obesity + othercirculatorysystem + otherrespsystem + renalfailure +
+ conditionrespiratoryarrest + sepsis + vascular + Northeast + West + South ,
+ data = covidcleansed)
summary(model.interactionH)

##
## Call:
## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + West + South, data = covidcleansed)
##
## Residuals:
##   Min    1Q Median    3Q   Max
## -3.949 -0.815 -0.048  0.728  3.349
## 
```

```

## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.4614   0.1684  14.62 < 2e-16 ***
## `age45-54`                  0.9397   0.0690  13.61 < 2e-16 ***
## adultrespdistSyn          -0.3738   0.2190  -1.71  0.08812 .
## allothcondsandcauses       0.4743   0.2043   2.32  0.02039 *
## alzheimer                 -3.1150   0.1956 -15.92 < 2e-16 ***
## cardiacArrest              -0.9263   0.2356  -3.93 8.9e-05 ***
## cardiacarrhythmia          -2.4140   0.2290 -10.54 < 2e-16 ***
## cerebrovascula             -2.3396   0.2421  -9.66 < 2e-16 ***
## chroniclr                  -2.0511   0.2302  -8.91 < 2e-16 ***
## covid19                     0.9267   0.2055   4.51 7.1e-06 ***
## diabetes                    -0.5683   0.2212  -2.57 0.01028 *
## heartfailure                -2.2222   0.2421  -9.18 < 2e-16 ***
## hypertensive                -0.8329   0.2276  -3.66 0.00026 ***
## influenza                   0.4008   0.2130   1.88 0.06007 .
## injuryPoisoning             -2.0530   0.2253  -9.11 < 2e-16 ***
## ischemicheart               -2.0256   0.2357  -8.59 < 2e-16 ***
## malignant                   -2.3342   0.2391  -9.76 < 2e-16 ***
## obesity                      -0.8505   0.2264  -3.76 0.00018 ***
## othercirculatorysystem      -1.4832   0.2385  -6.22 6.7e-10 ***
## otherrespsystem              -2.1648   0.2421  -8.94 < 2e-16 ***
## renalfailure                -1.2493   0.2255  -5.54 3.6e-08 ***
## conditionrespiratoryarrest -2.5858   0.2361 -10.95 < 2e-16 ***
## sepsis                       -0.9016   0.2253  -4.00 6.6e-05 ***
## vascular                     -3.0590   0.2017 -15.17 < 2e-16 ***
## Northeast                    0.1714   0.1077   1.59 0.11195
## West                         -0.0742   0.1004  -0.74 0.46034
## South                        0.5482   0.0960   5.71 1.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1334 degrees of freedom
## Multiple R-squared:  0.56,    Adjusted R-squared:  0.551
## F-statistic: 65.3 on 26 and 1334 DF,  p-value: <2e-16

```

## Final model

```

covidcleansed$`age45-54SQ` <- covidcleansed$`age45-54`^2
model.interactionH <- lm(log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn
+ allothcondsandcauses +
  alzheimer + cardiacArrest + cardiacarrhythmia + cerebrovascula +
  chroniclr + covid19 + diabetes + heartfailure + hypertensive +
  influenza + injuryPoisoning + ischemicheart + malignant +
  obesity + othercirculatorysystem + otherrespsystem + renalfailure +
  conditionrespiratoryarrest + sepsis + vascular + Northeast + West + South
, data = covidcleansed)
summary(model.interactionH)

##
## Call:

```

```

## lm(formula = log(num_deaths + 1) ~ `age45-54` + adultrespdistSyn +
##     allothcondsandcauses + alzheimer + cardiacArrest + cardiacarrhythmia +
##     cerebrovascula + chroniclr + covid19 + diabetes + heartfailure +
##     hypertensive + influenza + injuryPoisoning + ischemicheart +
##     malignant + obesity + othercirculatorysystem + otherrespsystem +
##     renalfailure + conditionrespiratoryarrest + sepsis + vascular +
##     Northeast + West + South, data = covidcleansed)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -3.949 -0.815 -0.048  0.728  3.349
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                2.4614   0.1684  14.62 < 2e-16 ***
## `age45-54`                  0.9397   0.0690  13.61 < 2e-16 ***
## adultrespdistSyn          -0.3738   0.2190  -1.71  0.08812 .
## allothcondsandcauses       0.4743   0.2043   2.32  0.02039 *  
## alzheimer                 -3.1150   0.1956 -15.92 < 2e-16 ***
## cardiacArrest              -0.9263   0.2356  -3.93 8.9e-05 ***
## cardiacarrhythmia          -2.4140   0.2290 -10.54 < 2e-16 ***
## cerebrovascula             -2.3396   0.2421  -9.66 < 2e-16 ***
## chroniclr                  -2.0511   0.2302  -8.91 < 2e-16 ***
## covid19                     0.9267   0.2055   4.51 7.1e-06 ***
## diabetes                   -0.5683   0.2212  -2.57 0.01028 *  
## heartfailure               -2.2222   0.2421  -9.18 < 2e-16 ***
## hypertensive                -0.8329   0.2276  -3.66 0.00026 *** 
## influenza                   0.4008   0.2130   1.88 0.06007 .  
## injuryPoisoning            -2.0530   0.2253  -9.11 < 2e-16 ***
## ischemicheart              -2.0256   0.2357  -8.59 < 2e-16 ***
## malignant                  -2.3342   0.2391  -9.76 < 2e-16 ***
## obesity                     -0.8505   0.2264  -3.76 0.00018 *** 
## othercirculatorysystem     -1.4832   0.2385  -6.22 6.7e-10 ***
## otherrespsystem             -2.1648   0.2421  -8.94 < 2e-16 ***
## renalfailure               -1.2493   0.2255  -5.54 3.6e-08 ***
## conditionrespiratoryarrest -2.5858   0.2361 -10.95 < 2e-16 ***
## sepsis                      -0.9016   0.2253  -4.00 6.6e-05 *** 
## vascular                   -3.0590   0.2017 -15.17 < 2e-16 ***
## Northeast                  0.1714   0.1077   1.59 0.11195
## West                        -0.0742   0.1004  -0.74 0.46034
## South                       0.5482   0.0960   5.71 1.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1334 degrees of freedom
## Multiple R-squared:  0.56, Adjusted R-squared:  0.551
## F-statistic: 65.3 on 26 and 1334 DF,  p-value: <2e-16

#plot(model.interactionH)

```

