Data Analysis Using Statistical Methods : Case Study of Covid Cases

KISLAY

Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ
Project Supervisor: Dr. Hadi Safari Katesari

ABSTRACT

This project aims to explore the relationship between hospitalizations and demographic variables such as age, gender, and race/ethnicity using data from different dates. We conducted exploratory data analysis, z-tests, ANOVA, categorical data analysis, regression analysis, resampling methods, and non-linear modeling to identify any patterns or trends in the data and test their significance. These findings can inform public health policies and interventions aimed at reducing hospitalizations and improving health equity.

1. INTRODUCTION

The COVID-19 pandemic has had a significant impact on public health worldwide, leading to millions of cases and deaths. In the United States, the pandemic has disproportionately affected certain demographic groups, such as older adults and racial/ethnic minorities, who have higher rates of hospitalizations and deaths. Understanding the factors that contribute to these disparities is crucial for developing effective prevention and treatment strategies.

In this project, we analyze data from different dates to investigate the relationship between hospitalizations and demographic variables such as age, gender, and race/ethnicity. We use a variety of statistical methods, including exploratory data analysis, t-tests, ANOVA, categorical data analysis, regression analysis, resampling methods, and non-linear modeling, to identify any patterns or trends in the data and test their significance.

Using these findings, we can inform public health policies and interventions aimed at reducing hospitalizations and improving health equity. For example, targeting interventions towards older age groups and racial/ethnic minorities may be more effective in reducing hospitalizations and deaths. The use of resampling methods and non-linear modeling can also improve the accuracy and generalizability of the models used to predict hospitalizations and inform public health decision-making.

2. DATA DESCRIPTION:

The data used in this analysis was collected from the following online resource:

https://catalog.data.gov/dataset/covid-19-daily-cases-deaths-and-hospitalizations. Read the data

```
data <- read.csv("/Users/kislaynandan/Downloads/COVID-19_Daily_Cases__Deaths__and_Hospitalizations.csv")
```

Structure of the data

```
str(data)
```

```
## ’data.frame’:    1117 obs. of  58 variables:
##  $ Date                                     : chr  "8/13/22" "6/16/22" "1/27/21" "7/22/22" ...
##  $ Cases...Total                            : int  450 749 657 825 686 450 351 913 440 970 ...
##  $ Deaths...Total                           : int  1 0 14 1 0 3 1 0 5 0 ...
##  $ Hospitalizations...Total                 : int  30 41 47 37 52 32 19 34 40 46 ...
##  $ Cases...Age.0.17                         : int  53 90 75 117 111 71 86 95 94 126 ...
##  $ Cases...Age.18.29                        : int  82 155 146 152 117 71 59 182 92 175 ...
##  $ Cases...Age.30.39                        : int  85 151 131 155 121 99 41 214 96 178 ...
##  $ Cases...Age.40.49                        : int  61 112 102 113 80 76 46 140 61 125 ...
##  $ Cases...Age.50.59                        : int  63 84 92 121 91 66 52 110 42 154 ...
##  $ Cases...Age.60.69                        : int  55 87 61 85 79 33 39 98 30 106 ...
##  $ Cases...Age.70.79                        : int  29 47 33 49 62 20 15 43 12 62 ...
##  $ Cases....Age.80.                         : int  22 23 17 32 25 14 13 31 13 44 ...
##  $ Cases...Age.Unknown                      : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ Cases...Female                           : int  259 455 337 482 401 259 193 514 245 551 ...
##  $ Cases...Male                             : int  188 291 320 342 281 190 157 396 195 419 ...
##  $ Cases...Unknown.Gender                   : int  3 3 0 1 4 1 1 3 0 0 ...
##  $ Cases...Latinx                           : int  118 148 201 191 192 98 106 214 72 230 ...
##  $ Cases...Asian.Non.Latinx                 : int  27 67 32 39 27 27 14 74 15 64 ...
##  $ Cases...Black.Non.Latinx                 : int  143 178 149 229 207 98 103 202 116 279 ...
##  $ Cases...White.Non.Latinx                 : int  85 215 174 223 156 130 85 221 147 219 ...
##  $ Cases...Other.Race.Non.Latinx            : int  17 34 29 35 25 29 17 51 33 51 ...
##  $ Cases...Unknown.Race.Ethnicity           : int  60 107 72 108 79 68 26 151 57 127 ...
##  $ Deaths...Age.0.17                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths...Age.18.29                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths...Age.30.39                       : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Deaths...Age.40.49                       : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Deaths...Age.50.59                       : int  0 0 0 0 0 1 1 0 1 0 ...
##  $ Deaths...Age.60.69                       : int  0 0 3 0 0 0 0 0 2 0 ...
##  $ Deaths...Age.70.79                       : int  1 0 4 1 0 0 0 0 0 0 ...
##  $ Deaths...Age.80.                         : int  0 0 7 0 0 1 0 0 1 0 ...
##  $ Deaths...Age.Unknown                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths...Female                          : int  1 0 4 1 0 3 0 0 4 0 ...
##  $ Deaths...Male                            : int  0 0 10 0 0 0 1 0 1 0 ...
##  $ Deaths...Unknown.Gender                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths...Latinx                          : int  0 0 2 0 0 0 1 0 1 0 ...
##  $ Deaths...Asian.Non.Latinx                : int  0 0 1 0 0 0 0 0 1 0 ...
##  $ Deaths...Black.Non.Latinx                : int  1 0 8 0 0 2 0 0 3 0 ...
##  $ Deaths...White.Non.Latinx                : int  0 0 3 1 0 1 0 0 0 0 ...
##  $ Deaths...Other.Race.Non.Latinx           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths...Unknown.Race.Ethnicity          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hospitalizations...Age.0.17              : int  0 4 0 1 3 4 0 2 1 3 ...
##  $ Hospitalizations...Age.18.29             : int  2 1 2 5 5 3 1 2 2 5 ...
##  $ Hospitalizations...Age.30.39             : int  1 8 4 2 6 2 0 3 6 3 ...
##  $ Hospitalizations...Age.40.49             : int  3 2 4 4 1 3 1 5 8 4 ...
##  $ Hospitalizations...Age.50.59             : int  5 2 8 4 10 4 3 4 8 8 ...
##  $ Hospitalizations...Age.60.69             : int  8 10 7 8 8 6 5 4 3 8 ...
##  $ Hospitalizations...Age.70.79             : int  3 5 14 5 10 4 3 4 4 8 ...
##  $ Hospitalizations...Age.80.               : int  8 9 8 8 9 6 6 10 8 7 ...
##  $ Hospitalizations...Age.Unknown           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hospitalizations...Female                 : int  15 24 27 18 31 11 9 24 23 22 ...
##  $ Hospitalizations...Male                   : int  15 17 20 19 21 21 10 10 17 24 ...
##  $ Hospitalizations...Unknown.Gender         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hospitalizations...Latinx                 : int  5 7 13 9 11 7 3 8 5 13 ...
##  $ Hospitalizations...Asian.Non.Latinx       : int  1 1 1 2 4 1 0 1 0 1 ...
##  $ Hospitalizations...Black.Non.Latinx       : int  13 17 16 16 24 14 11 13 25 19 ...
##  $ Hospitalizations...White.Non.Latinx       : int  8 12 14 7 11 7 4 8 8 11 ...
##  $ Hospitalizations...Other.Race.Non.Latinx : int  3 3 2 1 2 2 0 2 1 1 ...
##  $ Hospitalizations...Unknown.Race.Ethnicity: int  0 1 1 2 0 1 1 2 1 1 ...
```

The dataset includes information on daily cases, deaths, and hospitalizations related to COVID-19. The data is related to COVID-19 cases and deaths in a US for specific mentioned dates. The columns contain different metrics related to COVID-19 cases and deaths such as total number of cases and deaths, cases and deaths by age group, cases and deaths by gender, cases and deaths by ethnicity, and hospitalizations by age group and ethnicity.

In this analysis, descriptive statistics were performed on the data to better understand its distribution and characteristics. Figures 1-3 below show the distribution of total cases, deaths, and hospitalizations over time.

One issue with the data is that there are missing parameters. For example, the dataset includes a variable for cases by age, but there are cases labeled as "Age Unknown." Similarly, there are some cases labeled as "Unknown Gender" in the gender variable. These missing values may affect the accuracy of any statistical analysis conducted on the data. To address the issue of missing values, one solution would be to remove any rows with missing values.

Overall, this dataset provides valuable information on the impact of COVID-19 on various demographics. However, the missing values in some variables may limit the accuracy of any analysis conducted on the data.

3 Methodology

Z Test We have performed a two-sample z-test to compare the means of two number of covid cases of data: females and males. The null hypothesis is that there is no significant difference between the means of the two number covid cases, and the alternative hypothesis is that the means are not equal.

Linear Regression We have built a linear regression model to investigate the relationship between the total COVID cases and the cases in each age group.

We have used columns Cases…Total, which represents the total COVID cases, and the independent variables are the COVID cases in each age group.

The lm function is used to create the linear regression model, and the summary function is used to output the results of the model.

ANOVA Test We have performed an analysis of variance (ANOVA) on the relationship between the total number of COVID-19 cases and the cases in different age groups.

The hypothesis is whether there is a significant difference between the mean total COVID-19 cases across different age groups.

Categorical Data We have performed a chi-squared test to investigate the association between race and COVID-19 cases.

We created a sparse matrix datMat from the data, where each row corresponds to a specific date, and each column corresponds to a specific race group. The values in the matrix represent the number of COVID-19 cases for each race group on each date.

Resampling Methods We have performed bootstrap resampling to estimate the mean number of COVID-19 cases in each age group.

In this case, we are randomly sampling from the original data with replacement to generate multiple "bootstrap samples." For each bootstrap sample, we calculate the mean number of COVID-19 cases across all age groups.

By generating many bootstrap samples and calculating the mean number of cases in each, we can estimate the distribution of the sample mean and compute confidence intervals around it. We have generate 1000 bootstrap samples and computed the 2.5th and 97.5th percentiles of the bootstrap sample means as the lower and upper bounds of a 95% confidence interval.

Linear Model Selection and Regularization We have performed ridge regression on the given dataset of COVID cases in different age groups for specific dates.

4 Analysis and Result

Dimensions of the data

```
 dim(data)
```

```
## [1] 1117   58
```

```
attach(data)
```

Check Duplicates

```
options(max.print=2000)
duplicated(data)
```

```
##    [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [217] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [229] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [253] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [277] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [349] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [361] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [373] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [385] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [397] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [409] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [421] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [433] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [445] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [457] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [469] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [481] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [493] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [505] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [517] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [529] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [541] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [553] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [565] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [577] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [589] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [601] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [613] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [625] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [637] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [649] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [661] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [673] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [685] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [697] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [709] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [721] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [733] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [745] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [757] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [769] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [781] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [793] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [805] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [817] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [829] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [841] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [853] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [865] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [877] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [889] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [901] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [913] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [937] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [949] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [961] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [973] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [985] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [997] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1009] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1021] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1033] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1045] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1057] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1069] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1081] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1093] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1105] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1117] FALSE
```

- We have 0 duplicate values which is good.

Check missing values

```
 is.null(data)
```

```
## [1] FALSE
```

- We have 0 missing values which is very good.

Plots for Total Covid Cases

```
library(ggplot2)
library(dplyr)
```

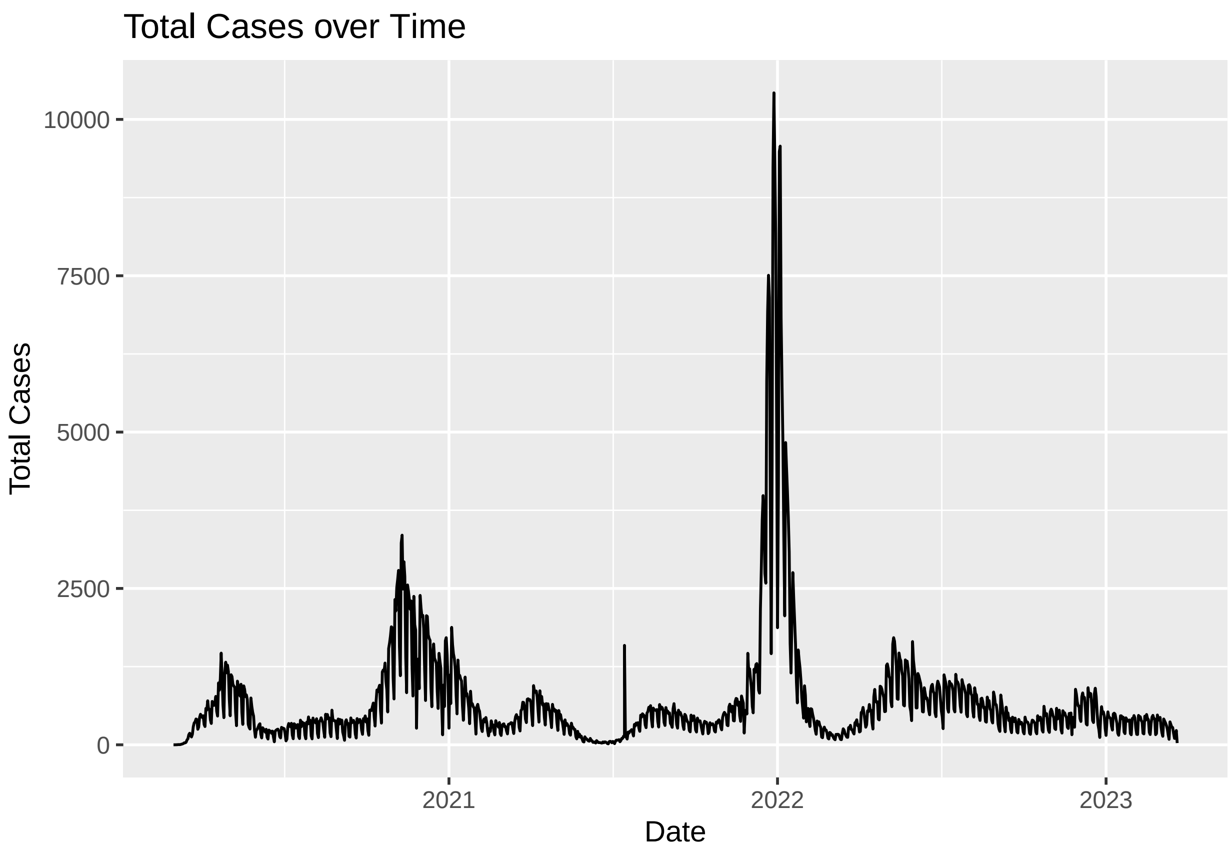
```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

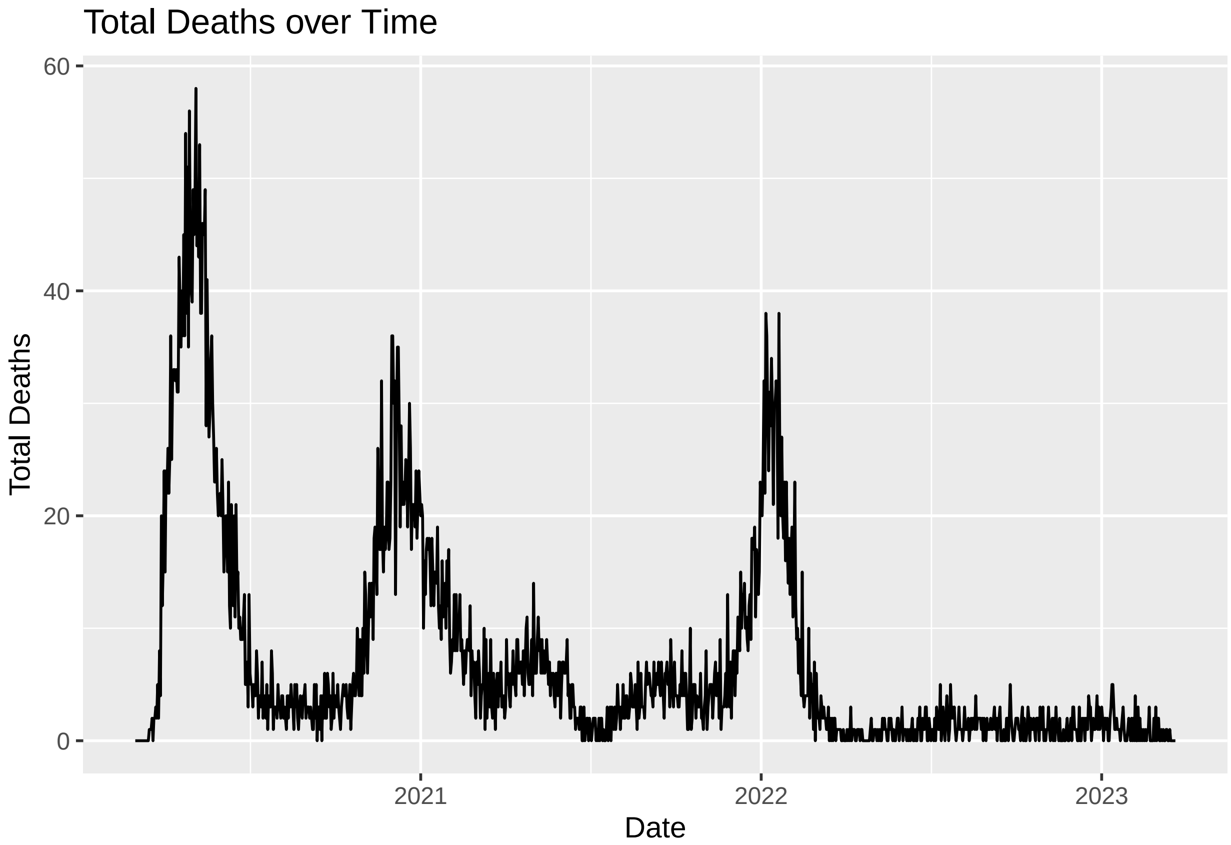
```
# Convert the date column to Date format
data$Date <- as.Date(data$Date, format = "%m/%d/%y")

# Create a line plot using geom_line()
# Create the plots
# Figure 1 - Total Cases
ggplot(data, aes(x = Date, y = `Cases...Total`)) +
  geom_line() +
  labs(title = "Total Cases over Time",
       x = "Date",
       y = "Total Cases")
```



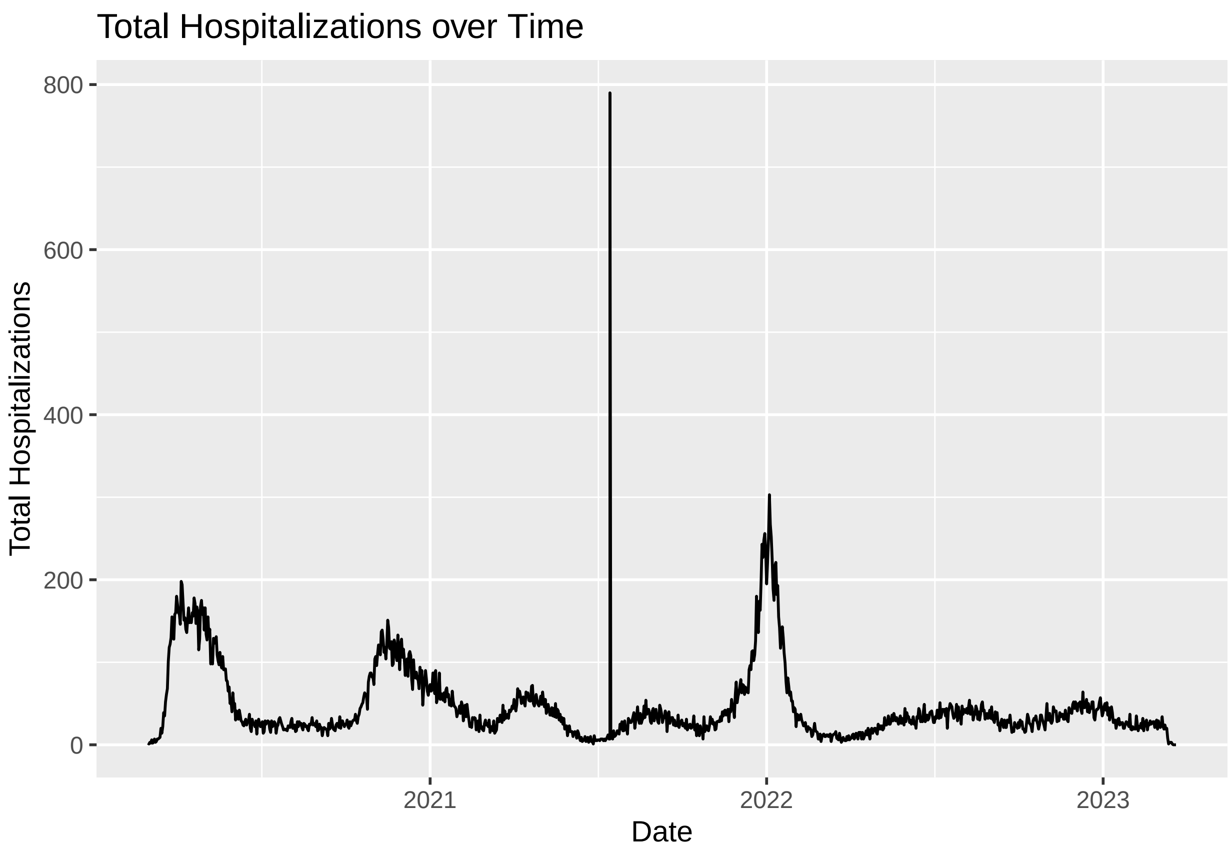
Plots for Total Deaths Cases

```
# Figure 2 - Total Deaths
ggplot(data, aes(x = Date, y = `Deaths...Total`)) +
  geom_line() +
  labs(title = "Total Deaths over Time",
       x = "Date",
       y = "Total Deaths")
```



Plots for Total Hospitalizations Cases

```
# Figure 3 - Total Hospitalizations
ggplot(data, aes(x = Date, y = `Hospitalizations...Total`)) +
  geom_line() +
  labs(title = "Total Hospitalizations over Time",
       x = "Date",
       y = "Total Hospitalizations")
```



4.1 Z Test

```
    # Extract the columns for females and males
females <- data$`Cases...Female`
males <- data$`Cases...Male`
```

```
# Calculate the sample means and standard deviations
mean_females <- mean(females)
mean_males <- mean(males)
sd_females <- sd(females)/sqrt(length(females))
sd_males <- sd(males)/sqrt(length(males))
```

```
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
#Perform Z test
z.test(females, males, alternative = "two.sided", sigma.x = 16.37, sigma.y = 13.49)
```

```
##
##  Two-sample z-Test
##
## data:  females and males
## z = 84.304, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  52.26275 54.75068
## sample estimates:
## mean of x mean of y
##  369.1638  315.6571
```

The output of the z-test shows that the test statistic z is 84.304 and the p-value is less than 2.2e-16, which is very small. This means that we can reject the null hypothesis and conclude that there is a significant difference between the means of the two populations.

The confidence interval shows that the true difference in means is likely to be between 52.26275 and 54.75068. Finally, the sample means of the two populations are also displayed. The mean of females is 369.1638 and the mean of males is 315.6571.

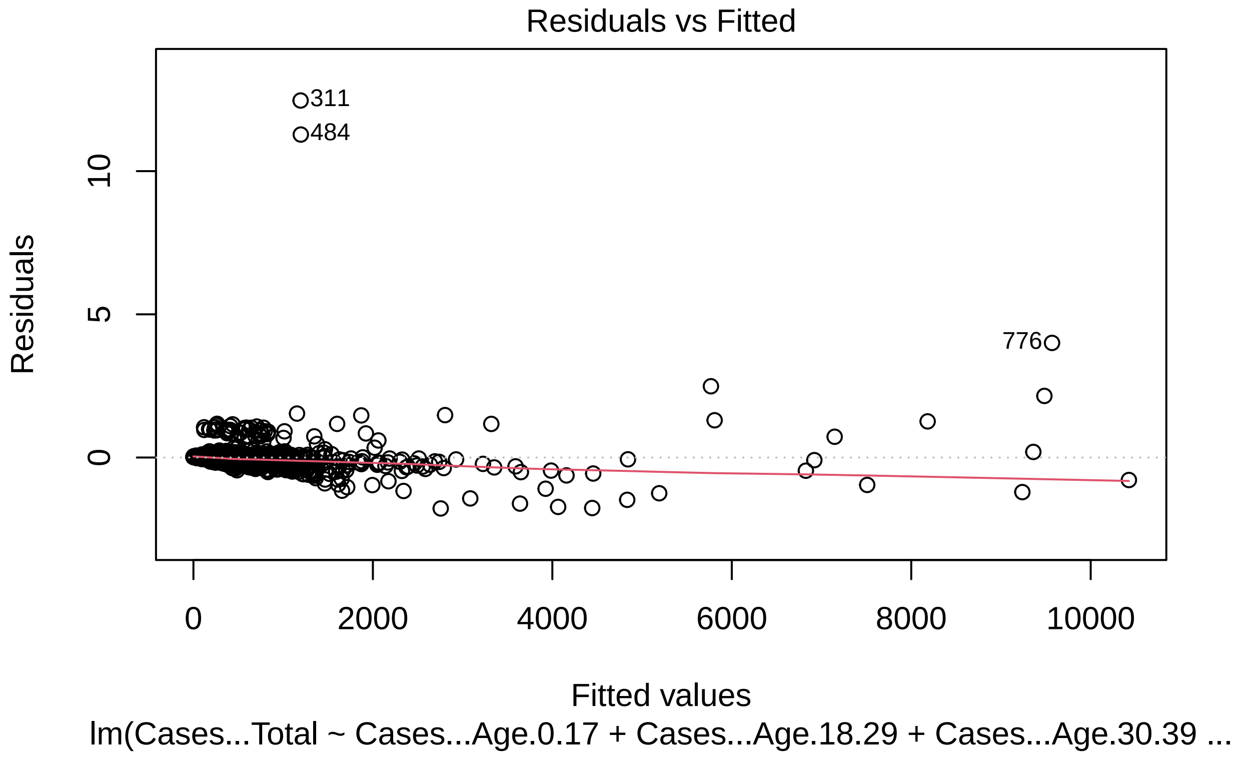
4.2 Linear Regression

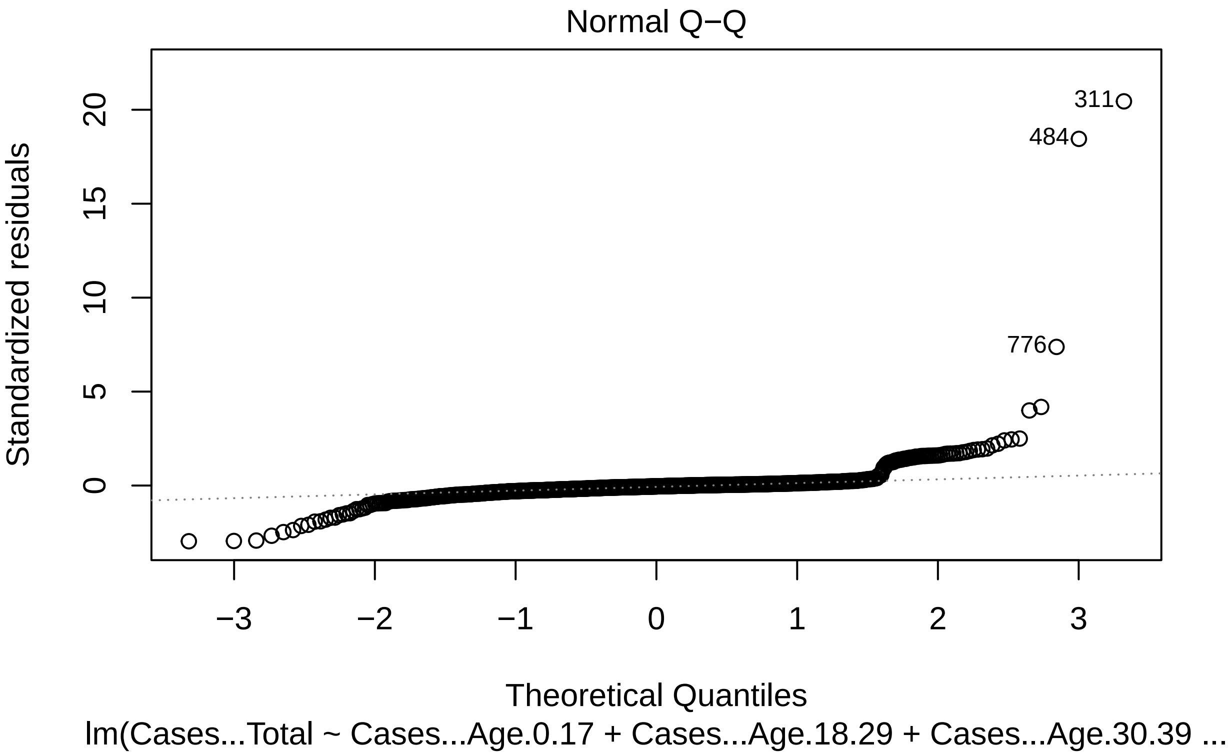
```
model2 <- lm(`Cases...Total` ~ `Cases...Age.0.17` + `Cases...Age.18.29` +
               `Cases...Age.30.39` + `Cases...Age.40.49` + `Cases...Age.50.59` +
               `Cases...Age.60.69` + `Cases...Age.70.79` + `Cases....Age.80.`,
             data = data)
summary(model2)
```

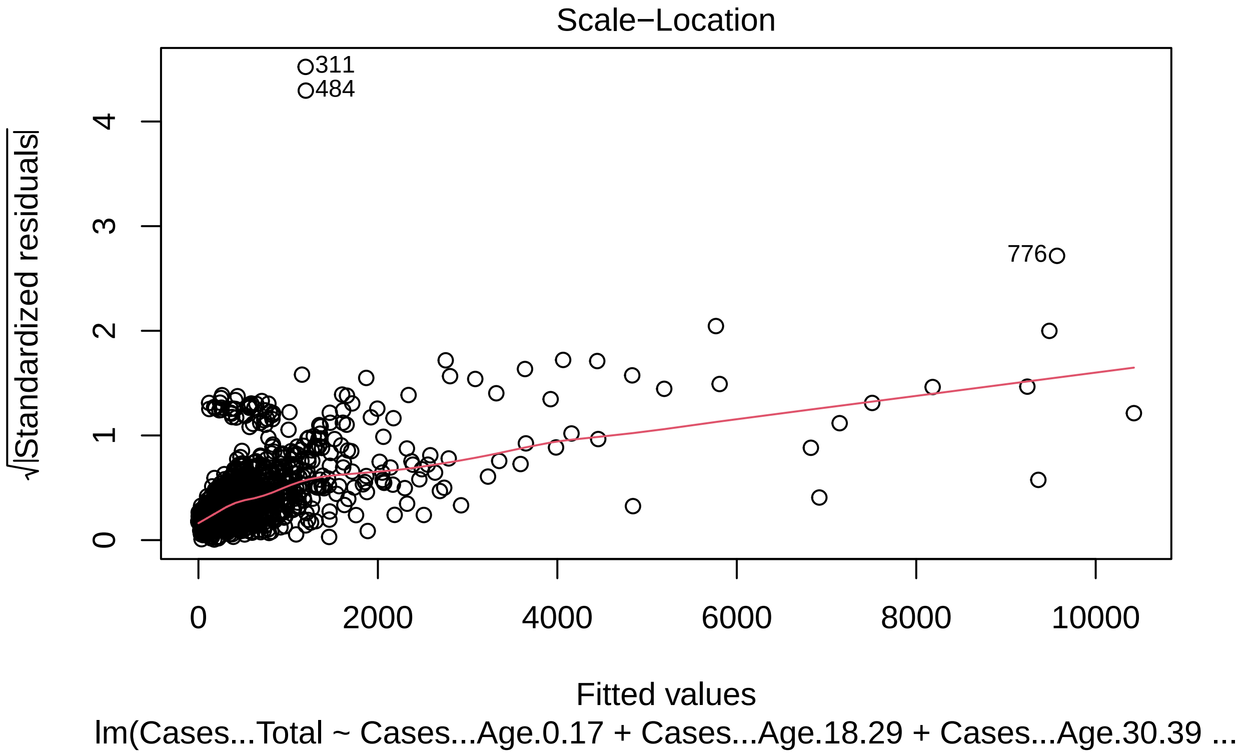
```
##
## Call:
## lm(formula = Cases...Total ~ Cases...Age.0.17 + Cases...Age.18.29 +
##     Cases...Age.30.39 + Cases...Age.40.49 + Cases...Age.50.59 +
##     Cases...Age.60.69 + Cases...Age.70.79 + Cases....Age.80.,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7782 -0.1266 -0.0276  0.0391 12.4665
##
## Coefficients:
##                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       -0.0200729  0.0255479   -0.786    0.432
## Cases...Age.0.17   1.0023511  0.0002235 4485.111   <2e-16 ***
## Cases...Age.18.29  0.9978679  0.0006962 1433.338   <2e-16 ***
## Cases...Age.30.39  1.0015375  0.0009715 1030.892   <2e-16 ***
## Cases...Age.40.49  0.9995928  0.0013654  732.087   <2e-16 ***
## Cases...Age.50.59  1.0021386  0.0014993  668.417   <2e-16 ***
## Cases...Age.60.69  1.0026614  0.0019572  512.286   <2e-16 ***
## Cases...Age.70.79  0.9872787  0.0025972  380.135   <2e-16 ***
## Cases....Age.80.   1.0043307  0.0027479  365.495   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.615 on 1108 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.684e+08 on 8 and 1108 DF,  p-value: < 2.2e-16
```

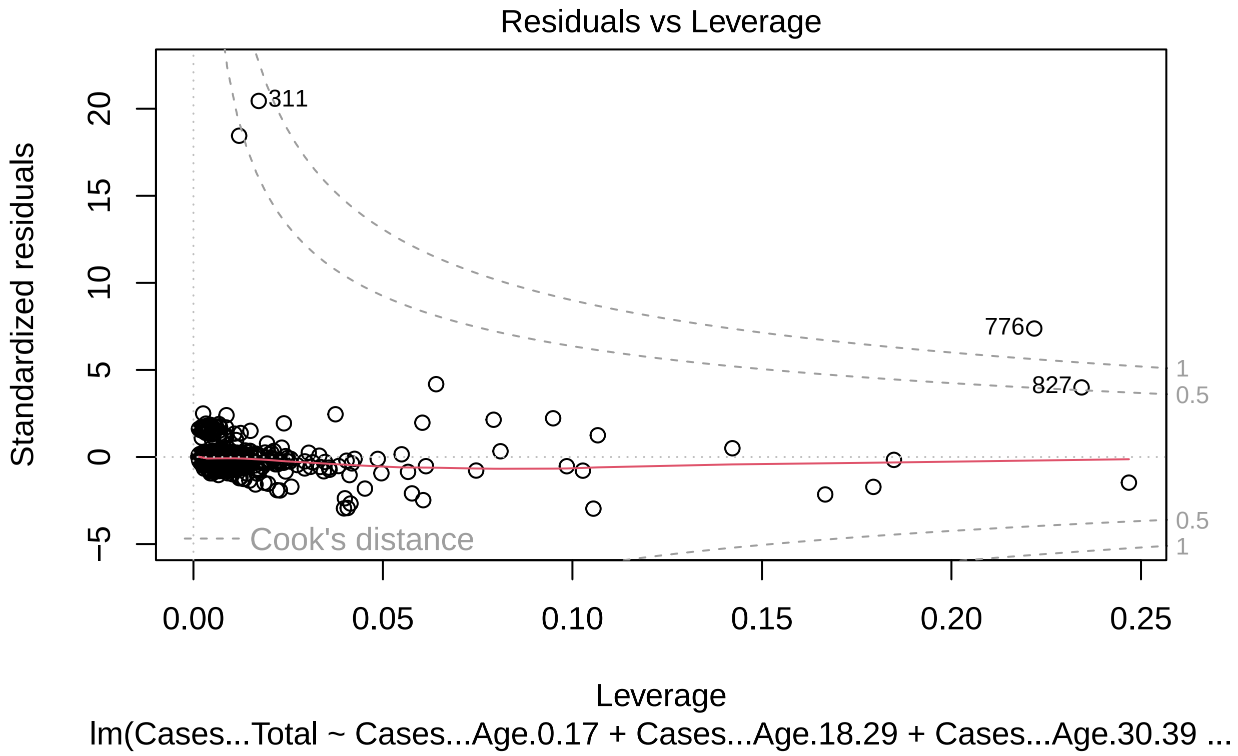
The p-value for each variable is less than 0.05, which suggests that each variable is significantly associated with the total COVID cases. The adjusted R-squared value of 1 indicates that the model explains all the variation in the dependent variable, and the F-statistic is significant (p-value < 2.2e-16), indicating that the overall model is significant.

In summary, the analysis suggests that the COVID cases in each age group are significantly associated with the total COVID cases, and the model can be used to predict the total COVID cases based on the cases in each age group.




Normal Q−Q
Standardized residuals
0
5
10
15
20
311
484
776
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Cases...Total ~ Cases...Age.0.17 + Cases...Age.18.29 + Cases...Age.30.39 ...


Scale−Location
311
484
776
√|Standardized residuals|
0
1
2
3
4
0
2000
4000
6000
8000
10000
Fitted values
lm(Cases...Total ~ Cases...Age.0.17 + Cases...Age.18.29 + Cases...Age.30.39 ...



4.3 ANOVA Test

```
model <- aov(`Cases...Total` ~ `Cases...Age.0.17` + `Cases...Age.18.29` +
              `Cases...Age.30.39` + `Cases...Age.40.49` + `Cases...Age.50.59` +
              `Cases...Age.60.69` + `Cases...Age.70.79` + `Cases....Age.80.`, data=data)
```

```
summary(model)
```

```
##                      Df    Sum Sq   Mean Sq   F value Pr(>F)
## Cases...Age.0.17      1 984649750 984649750 2.603e+09 <2e-16 ***
## Cases...Age.18.29     1 111841719 111841719 2.957e+08 <2e-16 ***
## Cases...Age.30.39     1   5828054   5828054 1.541e+07 <2e-16 ***
## Cases...Age.40.49     1   9740909   9740909 2.575e+07 <2e-16 ***
## Cases...Age.50.59     1   1755072   1755072 4.640e+06 <2e-16 ***
## Cases...Age.60.69     1    577222    577222 1.526e+06 <2e-16 ***
## Cases...Age.70.79     1    148736    148736 3.932e+05 <2e-16 ***
## Cases....Age.80.      1     50527     50527 1.336e+05 <2e-16 ***
## Residuals          1108       419         0
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The hypothesis being tested is whether there is a significant difference between the mean total COVID-19 cases across different age groups.

In this case, all age groups have very low p-values (2e-16), which means that there is a significant difference between the mean total COVID-19 cases across different age groups. This is further supported by the large F-values for each age group, indicating that the variance explained by the age group variable is much larger than the residual variance.

However, this doesn't tell us which groups are different from each other. We need to check individual testing that which test is a significant difference. Thus, we can perform Tukey's Test to determine exactly which group means are different.

4.4 Analysis of Categorical Data

```
library(Matrix)
datMat <- sparseMatrix(i = rep(1:length(data$Cases...Latinx), 6),
                       j = rep(1:6, each = length(data$Cases...Latinx)),
                       x = c(data$Cases...Latinx,
                             data$Cases...Asian.Non.Latinx,
                             data$Cases...Black.Non.Latinx,
                             data$Cases...White.Non.Latinx,
                             data$Cases...Other.Race.Non.Latinx,
                             data$Cases...Unknown.Race.Ethnicity))
```

```
rownames(datMat) <- names(data$Cases...Latinx)
colnames(datMat) <- c("Latinx", "Asian.Non.Latinx", "Black.Non.Latinx",
                      "White.Non.Latinx", "Other.Race.Non.Latinx",
                      "Unknown.Race.Ethnicity")

tableMat <- t(datMat) %*% datMat
```

```
chisq.test(tableMat)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tableMat
## X-squared = 983292409, df = 35, p-value < 2.2e-16
```

The test statistic is 983292409, with 35 degrees of freedom, and a p-value < 2.2e-16, which indicates strong evidence to reject the null hypothesis that the race groups are independent of COVID-19 cases.

Therefore, the null hypothesis in this case is that the observed counts of COVID cases in each race group are not significantly different from the expected counts based on the assumed probabilities. The alternative hypothesis would be that the observed counts are significantly different from the expected counts, suggesting that there may be a relationship between race and COVID cases.

This code is performing a chi-squared test of independence between race/ethnicity and COVID-19 cases.

First, it creates a sparse matrix called datMat where each row represents a COVID-19 case and each column represents a race/ethnicity category. The values in the matrix represent the number of cases in each category.

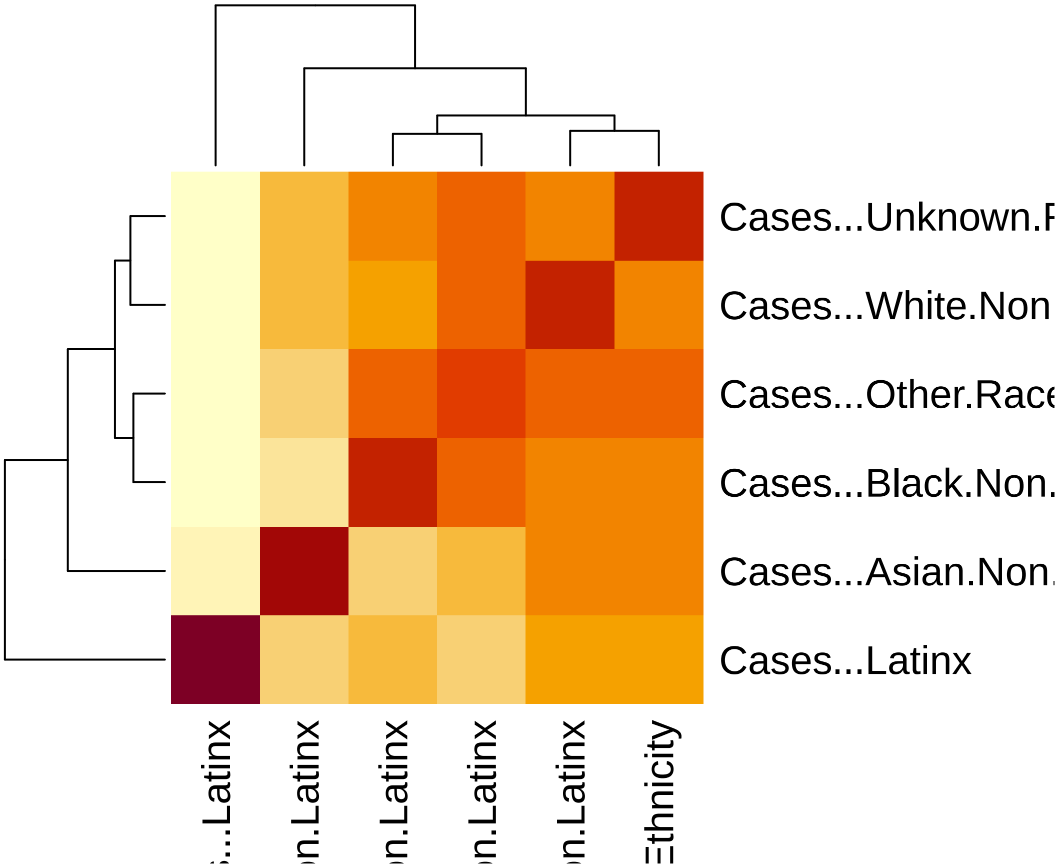
Next, it creates a table called tableMat by multiplying the transpose of datMat with datMat. This table shows the number of times each combination of race/ethnicity occurs in the dataset.

Finally, it performs a chi-squared test on tableMat to determine whether there is a significant association between race/ethnicity and COVID-19 cases

```
a <- cor(data[c("Cases...Latinx", "Cases...Asian.Non.Latinx", "Cases...Black.Non.Latinx",
            "Cases...White.Non.Latinx", "Cases...Other.Race.Non.Latinx",
            "Cases...Unknown.Race.Ethnicity")])
```

This code is calculating the correlation matrix between the columns "Cases. . . Latinx", "Cases. . . Asian.Non.Latinx", "Cases. . . Black.Non.Latinx", "Cases. . . White.Non.Latinx", "Cases. . . Other.Race.Non.Latinx", and "Cases. . . Unknown.Race.Ethnicity" in the data dataframe, and then creating a heatmap of the correlation matrix using the heatmap function.

The cor function computes the pairwise correlation between the columns of a dataframe, and returns a correlation matrix. The resulting matrix is then plotted using the heatmap function, which creates a heatmap where each cell is colored based on the correlation value, with a color scale on the side indicating the range of correlation values. This can help visualize the strength and direction of correlations between the different variables.



It has created a heatmap of the correlation matrix between the different Race group in the dataset. The heatmap visualizes the correlation coefficients between the variables, with warmer colors indicating higher correlation and cooler colors indicating lower correlation.

4.5 Resampling

```
library(dplyr)

bootstrap_mean_cases <- function(data) {
  sampled_data <- sample_n(data, nrow(data), replace = TRUE)
  selected_cols <- select(sampled_data, matches("^Cases\\.\\.\\.Age\\.|^Cases\\....Age\\."))
  mean_cases <- mean(unlist(selected_cols))
  return(mean_cases)
}
```

```
set.seed(123)
bootstrap_samples <- replicate(1000, bootstrap_mean_cases(data))
quantile(bootstrap_samples, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 70.12586 82.88374
```

We defined a function bootstrap_mean_cases() that generates a bootstrap sample from the input data and computes the mean of the selected columns with names matching the regular expression “Cases\.\.\.Age\.|Cases\....Age\.”.

Then, the function replicate() generates 1000 bootstrap samples by repeatedly calling bootstrap_mean_cases() and stores the computed means in bootstrap_samples. Finally, quantile() is used to compute the 95% confidence interval of the means using the 2.5th and 97.5th percentiles of bootstrap_samples.

We performed a bootstrap analysis to estimate the 95% confidence interval of the mean of the selected columns in data that contain the string “Cases…Age.” in their names.

4.6 Linear Model Selection and Regularization

```
library(glmnet)
```

```
## Loaded glmnet 4.1-7
```

```
x <- model.matrix(`Cases...Total` ~ `Cases...Age.0.17` + `Cases...Age.18.29` +
                   `Cases...Age.30.39` + `Cases...Age.40.49` + `Cases...Age.50.59` +
                   `Cases...Age.60.69` + `Cases...Age.70.79` + `Cases....Age.80.`,
                 data = data)[,-1]
y <- data$`Cases...Total`
```

```
cv.ridge <- cv.glmnet(x = x, y = y, alpha = 0,
         lambda = seq(0, 1, by = 0.01),
         type.measure = "mse",
         nfolds = 10)
```

```
cv.ridge
```

```
##
## Call:  cv.glmnet(x = x, y = y, lambda = seq(0, 1, by = 0.01), type.measure = "mse",      nfolds = 10
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure     SE Nonzero
## min   0.00   101  0.4146 0.1793       8
## 1se   0.44    57  0.5829 0.2117       8
```

```
best_lambda <- cv.ridge$lambda.min
ridge.model <- glmnet(x, y, alpha = 0, lambda = best_lambda)
summary(ridge.model)
```

```
##           Length Class     Mode
## a0        1      -none-    numeric
## beta      8      dgCMatrix S4
## df        1      -none-    numeric
## dim       2      -none-    numeric
## lambda    1      -none-    numeric
## dev.ratio 1      -none-    numeric
## nulldev   1      -none-    numeric
## npasses   1      -none-    numeric
## jerr      1      -none-    numeric
## offset    1      -none-    logical
## call      5      -none-    call
## nobs      1      -none-    numeric
```

```
ridge.model
```

```
##
## Call:  glmnet(x = x, y = y, alpha = 0, lambda = best_lambda)
##
##   Df %Dev Lambda
## 1  8  100      0
```

We are performing ridge regression on the COVID-19 data set to identify the most important predictors of the number of cases.

The output indicates that the ridge regression model has 8 degrees of freedom, 100% deviance explained (i.e., the model fits the data perfectly), and lambda value of 0.

4.7 Moving Beyond Linearity

```
library(splines)
library(mgcv)
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following objects are masked from 'package:BSDA':
##
##     Gasoline, Wheat
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
## This is mgcv 1.8-41. For overview type ’help("mgcv-package")’.
```

```
f <- Cases...Total ~ ns(`Cases...Age.0.17`) + ns(`Cases...Age.18.29`) +
              ns(`Cases...Age.30.39`) + ns(`Cases...Age.40.49`) +     ns(`Cases...Age.50.59`) +
              ns(`Cases...Age.60.69`) + ns(`Cases...Age.70.79`) + ns(`Cases....Age.80.`) + ns(`Cases...
```

```
model <- gam(f, data = data)
summary(model)
```

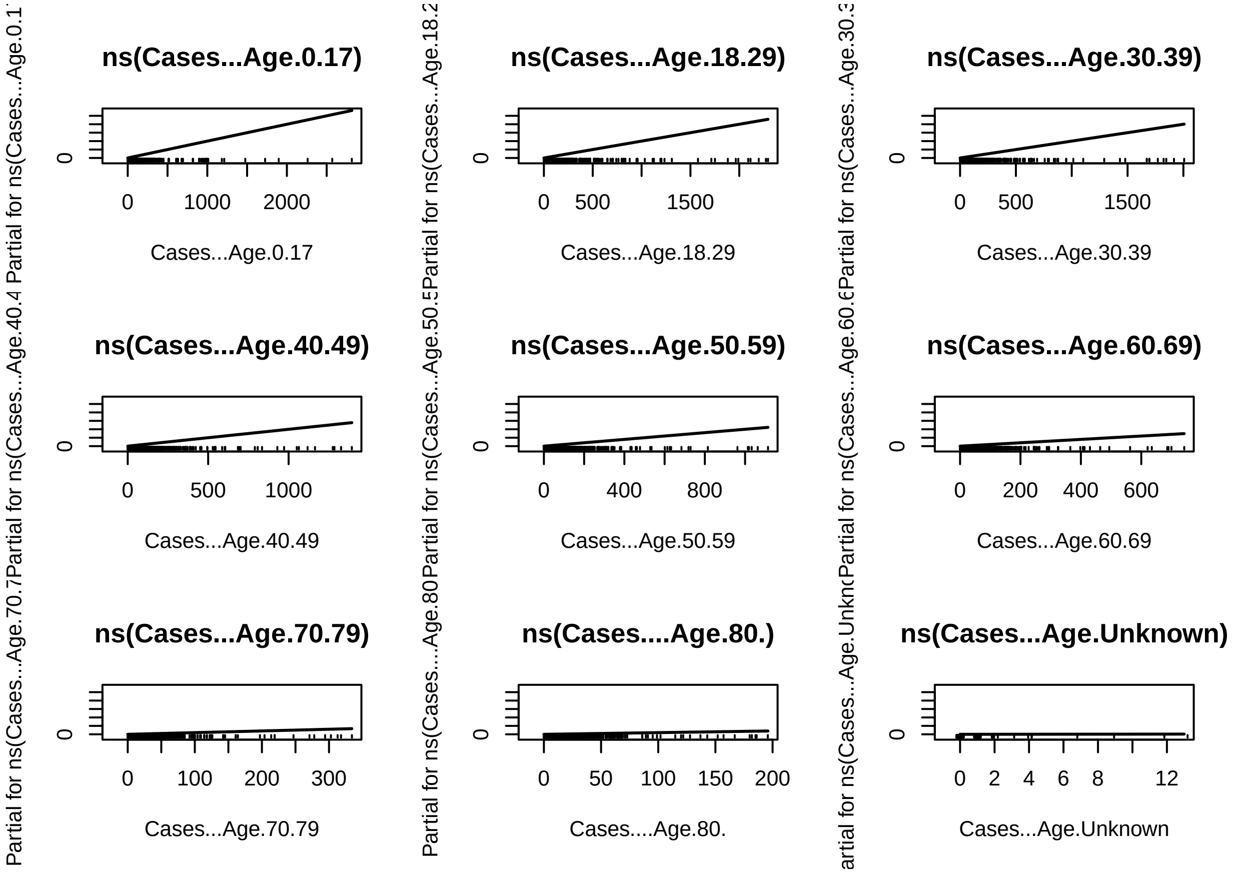
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Cases...Total ~ ns(Cases...Age.0.17) + ns(Cases...Age.18.29) +
##     ns(Cases...Age.30.39) + ns(Cases...Age.40.49) + ns(Cases...Age.50.59) +
##     ns(Cases...Age.60.69) + ns(Cases...Age.70.79) + ns(Cases....Age.80.) +
##     ns(Cases...Age.Unknown)
##
## Parametric coefficients:
##                           Estimate Std. Error  t value Pr(>|t|)
## (Intercept)              2.478e-11  1.928e-06        0        1
## ns(Cases...Age.0.17)     3.512e+03  6.212e-05 56540864   <2e-16 ***
## ns(Cases...Age.18.29)    2.861e+03  1.509e-04 18955204   <2e-16 ***
## ns(Cases...Age.30.39)    2.503e+03  1.837e-04 13625082   <2e-16 ***
## ns(Cases...Age.40.49)    1.737e+03  1.790e-04  9705225   <2e-16 ***
## ns(Cases...Age.50.59)    1.389e+03  1.573e-04  8830899   <2e-16 ***
## ns(Cases...Age.60.69)    9.254e+02  1.368e-04  6765179   <2e-16 ***
## ns(Cases...Age.70.79)    4.166e+02  8.252e-05  5048116   <2e-16 ***
## ns(Cases....Age.80.)     2.445e+02  5.075e-05  4817284   <2e-16 ***
## ns(Cases...Age.Unknown) 1.621e+01  3.676e-05   441116   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##
## R-sq.(adj) =      1   Deviance explained =   100%
## GCV = 2.1732e-09  Scale est. = 2.1537e-09  n = 1117
```

This code fits a generalized additive model (GAM) to the data using the gam() function from the mgcv package. The response variable is Cases...Total, and the predictor variables are the age groups Cases...Age.0.17, Cases...Age.18.29, Cases...Age.30.39, Cases...Age.40.49, Cases...Age.50.59, Cases...Age.60.69, Cases...Age.70.79, Cases....Age.80., and Cases...Age.Unknown.

The predictor variables are transformed using natural splines with the ns() function from the splines package. The default number of degrees of freedom is used for each spline, which is determined by the mgcv package based on the data.

The summary() function is used to print a summary of the fitted model, including estimates of the smoothing parameters and standard errors, degrees of freedom for each term, and the deviance explained by the model.

```
plot(model, pages = 1, all.terms = TRUE)
```



Each plot shows the partial effect of the corresponding spline term on the response variable, while holding all other predictors constant. The x-axis represents the range of the predictor variable, while the y-axis represents the partial effect of the spline term on the response variable.

In this model, we have taken nine spline terms, one for each age group. Therefore, the nine plots show the partial effects of each age group on the total number of cases, while holding all other age groups constant.

The plots can be useful for visualizing how the relationship between age and cases changes across different age groups.

Conclusion

Z test - We can reject the null hypothesis and conclude that there is a significant difference between the means of the two populations.

Linear Regression - In summary, the analysis suggests that the COVID cases in each age group are significantly associated with the total COVID cases, and the model can be used to predict the total COVID cases based on the cases in each age group.

ANOVA Test - We can reject the null hypothesis and conclude that there is a significant difference between the mean number of COVID-19 cases in each age group.

Chi-square test gives relationships for categorical values of data.

Linear Regression gives a good r2 score of 1.

The output shows that the range of the means of the bootstrap samples is between 70.12586 and 82.88374.

References

- https://catalog.data.gov/dataset/covid-19-daily-cases-deaths-and-hospitalizations.
- https://www.analystsoft.com/en/products/statplus/content/help/analysis_analysis_of_varianceone_way_anova/
- https://www.scribbr.com/statistics/anova-in-r/
- https://statsandr.com/blog/chi-square-test-of-independence-in-r/