DATA 606 Final Project

Fomba Kassoh & Souleymane Doumbia

2023-11-19

Contents

Abstract

The goal of this analysis was to determine the most appropriate regression model for COVID-19 case data, specifically looking at the daily counts of cases, hospitalizations, and deaths for the years 2021 and 2022. The data exhibited characteristics typical of count data, including right-skewness and overdispersion, challenging the assumptions of standard linear regression models.

Initial diagnostics using histograms, Q-Q plots, and boxplots revealed that the distributions of the variables were not normally distributed and contained outliers. The Q-Q plots, in particular, showed significant deviations from the theoretical quantiles of a normal distribution, indicating heavy-tailed distributions for all three variables. Scatter plots suggested nonlinear relationships and potential associations among the variables.

Given these preliminary findings, a standard linear regression model was deemed unsuitable due to the non-normality of residuals and heteroscedasticity. Consequently, we explored Generalized Linear Models (GLMs) with different distributions that are more robust to the peculiarities of count data.

A Poisson GLM was initially considered, as it is a common choice for modeling count data. However, diagnostic checks indicated significant overdispersion and the presence of outliers, which the Poisson model inherently cannot handle, as it assumes the mean and variance of the data to be equal.

To address the overdispersion, a Negative Binomial GLM was fitted, which introduces an additional parameter to account for the variance exceeding the mean. Subsequent diagnostic plots revealed an improved fit over the Poisson model, with reduced issues of overdispersion and fewer influential outliers, as evidenced by Cook's distance.

The analyses suggest that for count data exhibiting overdispersion and skewness, as is often the case in epidemiological data, Negative Binomial regression provides a better fit than both linear and Poisson regression models. This model adequately addresses the variance structure of the data, providing more reliable estimates for the effects of hospitalizations and cases on COVID-19-related deaths. For accurate modeling of such data, it is crucial to consider the underlying distribution and employ a model that can capture its nuances.

Overview

Context of Data Collection

The data used in this analysis is comprised of Covid-19 death and hospitalization counts. The data was retrieved from the NYC OpenData portal.The data can be view by following the following link. https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3. The COVID-19 Daily Counts of Cases, Hospitalizations, and Death are observed by the NYS Department of Health and are publicly available. The data can be retried through an API end point or by downloading a csv or json format files.We are using the following API end point. https://data.cityofnewyork.us/resource/rc75-m7u3.csv. The following is the context of the Covid-19 data collection:

Early Pandemic (Early 2020): Limited testing led to potential under reporting. **Expanded Data Collection (Mid-2020):** By mid-2020, testing became more widespread, allowing for more comprehensive data collection, including test positivity rates and detailed demographic information. **Vaccination Rollout (Late 2020/Early 2021):** Vaccination data started being collected around December 2020 to early 2021, as COVID-19 vaccines were approved and began to be administered worldwide. **Impact of Public Health Campaigns (Throughout 2020 and 2021):** Throughout 2020 and into 2021, public health campaigns, particularly for vaccination and preventive measures, likely influenced the trends observed in the data. **Adaptation to Variants (2021 Onwards):** Starting in 2021, with the emergence of variants like Delta and later Omicron, continuous monitoring became crucial to assess vaccine effectiveness against these new strains. **Long-Term Monitoring (2021 Onwards):** As the pandemic progressed into 2021 and beyond,

the focus shifted to long-term effects of COVID-19, enduring vaccine efficacy, and evaluating the impact of different public health policies.

What are the variables in the data

The variables in the data as show below

Load the data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
# Loading the dataset using the provided API
file_path <- 'https://data.cityofnewyork.us/resource/rc75-m7u3.csv'
covid_data <- read_csv(file_path)
```

```
## Rows: 1000 Columns: 67
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (66): case_count, probable_case_count, hospitalized_count, death_count,...
## dttm  (1): date_of_interest
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Displaying the first few rows of the dataset to understand its structure
suppressWarnings({
  glimpse(covid_data)
})
```

```
## Rows: 1,000
## Columns: 67
## $ date_of_interest               <dttm> 2020-02-29, 2020-03-01, 2020-03-02, 2~
## $ case_count                     <dbl> 1, 0, 0, 1, 5, 3, 8, 7, 21, 57, 69, 15~
## $ probable_case_count            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ hospitalized_count             <dbl> 1, 1, 2, 7, 2, 14, 8, 8, 18, 36, 60, 7~
## $ death_count                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ probable_death_count           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ case_count_7day_avg            <dbl> 0, 0, 0, 0, 0, 0, 3, 3, 6, 15, 24, 46,~
## $ all_case_count_7day_avg        <dbl> 0, 0, 0, 0, 0, 0, 3, 3, 6, 15, 24, 46,~
## $ hosp_count_7day_avg            <dbl> 0, 0, 0, 0, 0, 0, 5, 6, 8, 13, 21, 32,~
```

```
## $ death_count_7day_avg            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ all_death_count_7day_avg        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bx_case_count                   <dbl> 0, 0, 0, 0, 0, 0, 2, 0, 3, 4, 8, 19, 2~
## $ bx_probable_case_count          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bx_hospitalized_count           <dbl> 0, 1, 0, 1, 0, 1, 1, 1, 5, 7, 7, 23, 1~
## $ bx_death_count                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bx_probable_death_count         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bx_case_count_7day_avg          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 5, 9,~
## $ bx_probable_case_count_7day_avg <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bx_all_case_count_7day_avg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 5, 9,~
## $ bx_hospitalized_count_7day_avg  <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 3, 6, 9,~
## $ bx_death_count_7day_avg         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bx_all_death_count_7day_avg     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bk_case_count                   <dbl> 0, 0, 0, 0, 1, 3, 1, 2, 5, 16, 11, 31,~
## $ bk_probable_case_count          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bk_hospitalized_count           <dbl> 1, 0, 2, 3, 1, 3, 1, 3, 8, 11, 13, 11,~
## $ bk_death_count                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bk_probable_death_count         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bk_case_count_7day_avg          <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 2, 4, 6, 10, 2~
## $ bk_probable_case_count_7day_avg <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bk_all_case_count_7day_avg      <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 2, 4, 6, 10, 2~
## $ bk_hospitalized_count_7day_avg  <dbl> 0, 0, 0, 0, 0, 0, 2, 2, 3, 4, 6, 7, 11~
## $ bk_death_count_7day_avg         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ bk_all_death_count_7day_avg     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ mn_case_count                   <dbl> 1, 0, 0, 0, 2, 0, 3, 1, 6, 24, 24, 62,~
## $ mn_probable_case_count          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ mn_hospitalized_count           <dbl> 0, 0, 0, 1, 1, 5, 3, 0, 1, 9, 12, 19, ~
## $ mn_death_count                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ mn_probable_death_count         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ mn_case_count_7day_avg          <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 2, 5, 9, 17, 3~
## $ mn_probable_case_count_7day_avg <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ mn_all_case_count_7day_avg      <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 2, 5, 9, 17, 3~
## $ mn_hospitalized_count_7day_avg  <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 2, 3, 4, 7, 9,~
## $ mn_death_count_7day_avg         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ mn_all_death_count_7day_avg     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ qn_case_count                   <dbl> 0, 0, 0, 1, 2, 0, 1, 3, 6, 10, 24, 40,~
## $ qn_probable_case_count          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ qn_hospitalized_count           <dbl> 0, 0, 0, 2, 0, 4, 2, 4, 4, 8, 23, 23, ~
## $ qn_death_count                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ qn_probable_death_count         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ qn_case_count_7day_avg          <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 2, 3, 7, 12, 2~
## $ qn_probable_case_count_7day_avg <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ qn_all_case_count_7day_avg      <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 2, 3, 7, 12, 2~
## $ qn_hospitalized_count_7day_avg  <dbl> 0, 0, 0, 0, 0, 0, 1, 2, 2, 3, 6, 10, 1~
## $ qn_death_count_7day_avg         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ qn_all_death_count_7day_avg     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ si_case_count                   <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 1, 3, 2, 3, 13~
## $ si_probable_case_count          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ si_hospitalized_count           <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 0, 2, 5, 2, 3,~
## $ si_death_count                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ si_probable_death_count         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ si_probable_case_count_7day_avg <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ si_case_count_7day_avg          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 3,~
## $ si_all_case_count_7day_avg      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 3,~
```

```
## $ si_hospitalized_count_7day_avg  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2, 2,~
## $ si_death_count_7day_avg         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ si_all_death_count_7day_avg     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ incomplete                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

Dependent Variable (death_count):

This is the outcome or response variable that we are trying to predict or explain. In the context of COVID-19, "Death Count" represents the number of deaths attributed to the virus on a given day. Independent Variables:

Description of Independent Variables

We are using the following two independent variable for our analysis:

hospitalized_count This variable represents the number of people hospitalized due to COVID-19 on a given day. It's an independent variable because it is presumed to influence or explain the variation in the death count.

case_count This is the total number of new confirmed COVID-19 cases. As with hospitalizations, this variable is used to predict or explain changes in the death count.

Research Question

"How do the number of COVID-19 hospitalizations and case counts predict or influence the number of COVID-19 related deaths?"

Summary Statistics

Based on the context of data collection, 2020 does not include vaccination and wide scale data collection. We have therefore excluded it from the analysis so that the condition/context under which data was collected is similar. The data set from the portal does not include 2023. We are there using the data for 2021 and 2022 for our analysis.

Determine if there are missing values in any column

Base on the search for missing values below, there are no missing values.

```
# Use dplyr and tidyr to find missing data
covid_data %>%
  summarise_all(~ sum(is.na(.)))
```

```
## # A tibble: 1 x 67
##   date_of_interest case_count probable_case_count hospitalized_count death_count
##              <int>      <int>               <int>              <int>       <int>
## 1                0          0                   0                  0           0
## # i 62 more variables: probable_death_count <int>, case_count_7day_avg <int>,
```

```
## #   all_case_count_7day_avg <int>, hosp_count_7day_avg <int>,
## #   death_count_7day_avg <int>, all_death_count_7day_avg <int>,
## #   bx_case_count <int>, bx_probable_case_count <int>,
## #   bx_hospitalized_count <int>, bx_death_count <int>,
## #   bx_probable_death_count <int>, bx_case_count_7day_avg <int>,
## #   bx_probable_case_count_7day_avg <int>, ...
```

Summary Statistics

The summary statistics below shows considerable variability in daily cases, hospitalizations, and deaths, likely reflecting the waves and changing dynamics of the COVID-19 pandemic during 2021 and 2022. The median of all three metrics (cases, hospitalizations, deaths) lower than the mean, pointing to a right-skewed distribution.The right-skewed distributions suggest that most days were on the lower end of the spectrum, but there were periods with significantly higher counts. The wide range between the minimum and maximum values for each metric highlights the fluctuating intensity of the pandemic over these years.

```r
library(dplyr)

# Convert 'date_of_interest' to Date and filter the data for 2021 and 2022
covid_data_2021_2022 <- covid_data %>%
  mutate(date_of_interest = as.Date(date_of_interest)) %>%
  filter(format(date_of_interest, "%Y") %in% c("2021", "2022"))

# Check if the required columns exist
if(all(c("case_count", "hospitalized_count", "death_count") %in% names(covid_data_2021_2022))) {
  # Use dplyr::select to avoid namespace conflict
  relevant_data_2021_2022 <- dplyr::select(covid_data_2021_2022, case_count, hospitalized_count, death_

  # Generate summary statistics for the selected columns (2021 and 2022 data)
  summary_statistics <- relevant_data_2021_2022 %>%
    summarise(across(c(case_count, hospitalized_count, death_count),
                     list(mean = ~ mean(., na.rm = TRUE),
                          median = ~ median(., na.rm = TRUE),
                          sd = ~ sd(., na.rm = TRUE),
                          min = ~ min(., na.rm = TRUE),
                          max = ~ max(., na.rm = TRUE),
                          IQR = ~ IQR(., na.rm = TRUE),
                          quantile_25 = ~ quantile(., probs = 0.25, na.rm = TRUE),
                          quantile_75 = ~ quantile(., probs = 0.75, na.rm = TRUE)
                     ))) %>%
    pivot_longer(cols = everything(), names_to = c(".value", "statistic"), names_pattern = "(case_count

  # Display the summary statistics
  print(summary_statistics)
} else {
  stop("One or more required columns are missing from the dataset.")
}
```

```
## # A tibble: 8 x 4
##   statistic   case_count hospitalized_count death_count
##   <chr>            <dbl>              <dbl>       <dbl>
## 1 mean             3073.               156.        23.9
## 2 median           1724                107         12
```

```
## 3 sd               6122.              169.        26.5
## 4 min                96                13          1
## 5 max             55009              1289        131
## 6 IQR              2007               112         23
## 7 quantile_25       944                56          7
## 8 quantile_75      2951               168         30
```

Appropiate Visualizations

Histogram of Death Count (Left Plot):

The histogram for all three metrics (cases, hospitalizations, deaths) shows a right-skewed distribution, indicating that most of the days had lower case, hospitalization, and death counts. The peak of the distributions is towards the lower end, suggesting that on many days, the counts were relatively low. The long tail to the right indicates that there were days with significantly higher counts, although these were less frequent.

Box Plot of Death Count (Right Plot):

The box plot further illustrates the skewed nature of the data. The median (indicated by the line inside the box) is closer to the lower quartile, which aligns with the histogram showing a concentration of data on the lower end. The presence of several points above the upper whisker indicates outliers, representing days with unusually high death counts.The interquartile range (the box) is relatively small compared to the range of the entire dataset, underscoring the skewness and the presence of extreme values on certain days.

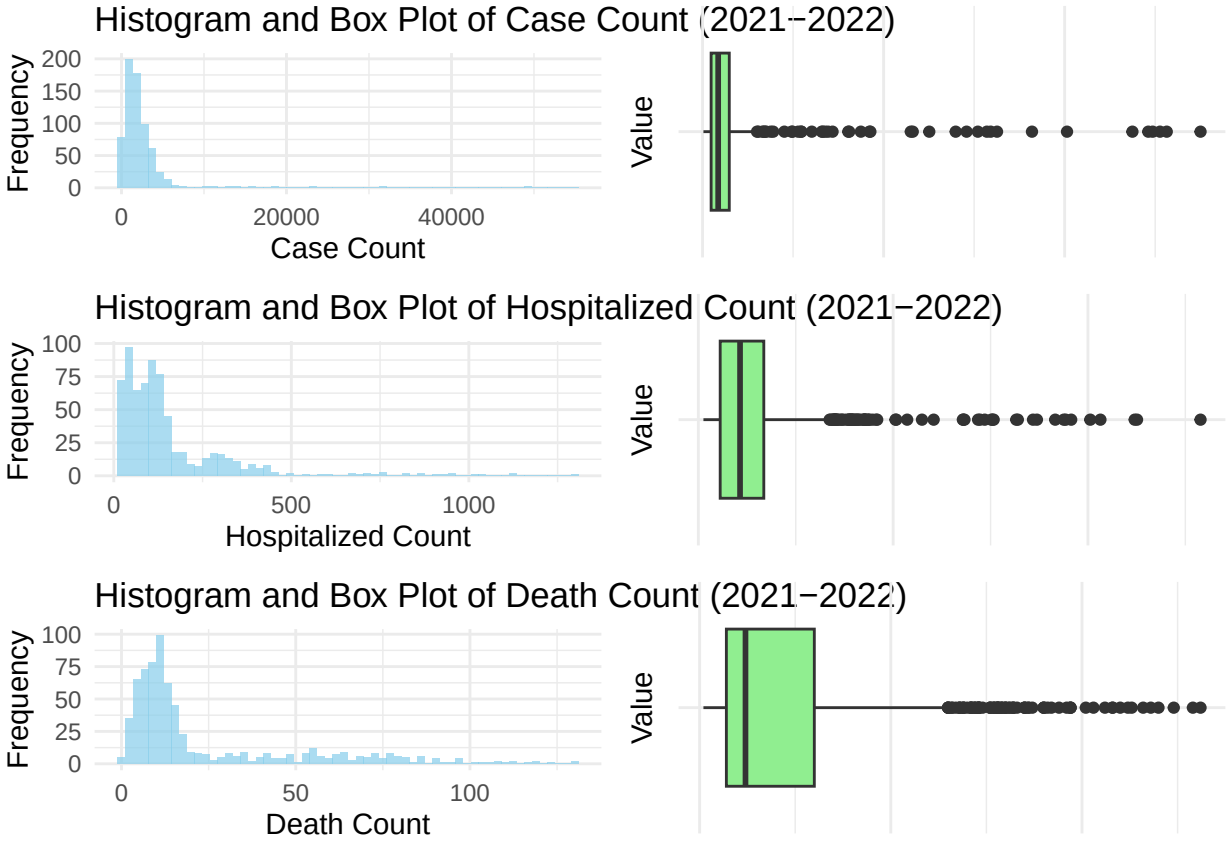
```r
library(dplyr)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

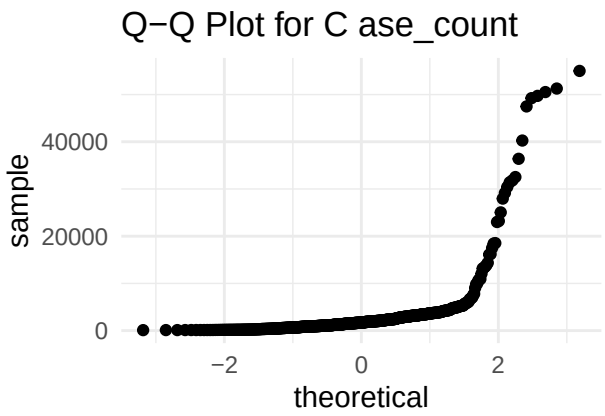
```r
# Function to create combined histogram and box plot
combined_plot <- function(data, column_name, title, xlabel) {
  numeric_data <- data[[column_name]]

  histogram_plot <- ggplot(data.frame(x = numeric_data), aes(x = x)) +
                    geom_histogram(bins = 60, fill = 'skyblue', alpha = 0.7) +
                    labs(title = title, x = xlabel, y = 'Frequency') +
                    theme_minimal()

  boxplot_plot <- ggplot(data.frame(x = numeric_data), aes(x = x, y = '')) +
                  geom_boxplot(fill = 'lightgreen') +
                  labs(y = 'Value') +
                  theme_minimal() +
                  theme(axis.title.x = element_blank(),
                        axis.text.x = element_blank(),
```

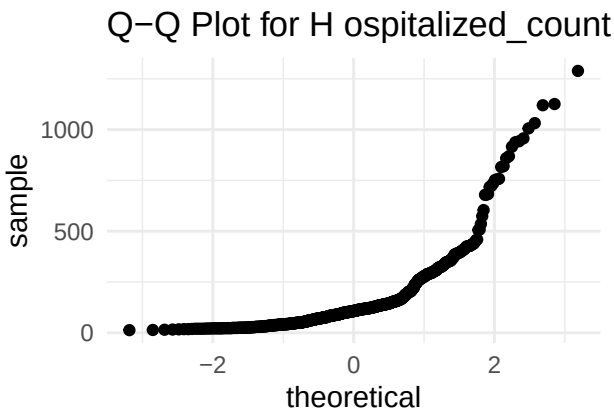
```
                        axis.ticks.x = element_blank())

  list(histogram_plot, boxplot_plot)
}

# Generate plots for each column
case_count_plots <- combined_plot(covid_data_2021_2022, 'case_count', 'Histogram and Box Plot of Case C
hospitalized_count_plots <- combined_plot(covid_data_2021_2022, 'hospitalized_count', 'Histogram and Bo
death_count_plots <- combined_plot(covid_data_2021_2022, 'death_count', 'Histogram and Box Plot of Deat

# Arrange all plots in a grid layout
grid.arrange(grobs = c(case_count_plots, hospitalized_count_plots, death_count_plots), ncol = 2)
```

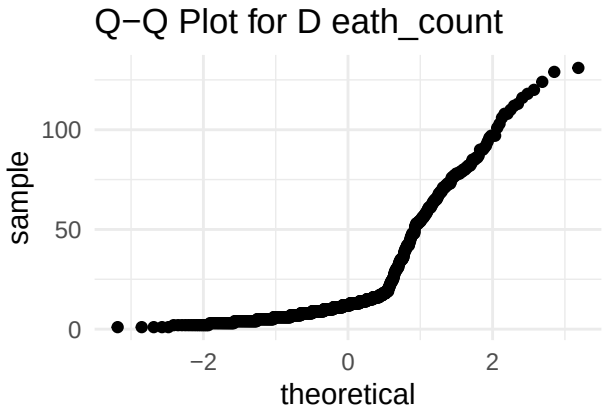



```
library(ggplot2)
library(gridExtra)


# Convert date_of_interest to a Date or year format if it's not already
covid_data$date_of_interest <- as.Date(covid_data$date_of_interest)

# Filter for 2021 and 2022
covid_data_2021_2022 <- covid_data[format(covid_data$date_of_interest, "%Y") %in% c("2021", "2022"), ]

# Function to create a Q-Q plot using ggplot2
create_qq_plot <- function(data, variable) {
```

```
    data_vector <- na.omit(data[[variable]])
    ggplot(data.frame(data_vector), aes(sample = data_vector)) +
        stat_qq() +
        ggtitle(paste('Q-Q Plot for', toupper(substring(variable, 1, 1)), substring(variable, 2))) +
        theme_minimal()

}

# Create plots for each variable
p1 <- create_qq_plot(covid_data_2021_2022, 'case_count')
p2 <- create_qq_plot(covid_data_2021_2022, 'hospitalized_count')
p3 <- create_qq_plot(covid_data_2021_2022, 'death_count')

grid.arrange(p1, p2, p3, ncol = 2)
```








```
# You can view them individually by just typing their names in the console
```

Scatter plots for the filtered dataset (2021-2022)

Death Count vs. Case Count The plot shows a non-linear relationship between the case count and death count. There appears to be a positive association; as case counts increase, death counts also tend to increase. However, the relationship does not seem to be linear, especially at higher case counts where the increase in death count is not proportional. There is a clear pattern indicating possible exponential or polynomial growth.

Death Count vs. Hospitalized Count This plot also indicates a positive relationship between hospitalized count and death count, with a similar non-linear pattern. As with case count, the increase in death count does not seem linearly proportional to the increase in hospitalized count, especially at higher values.

Appropriate Regression Model Given the non-linear patterns observed in both scatter plots, a linear regression model may not provide the best fit for the data. Possible approaches to modeling these relationships could include

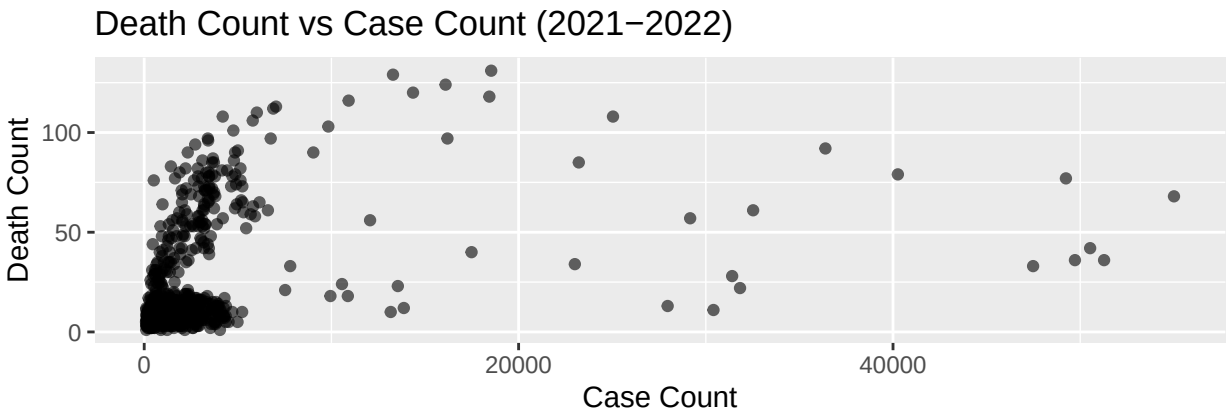
1. Polynomial Regression: To capture the curvature in the relationship, a polynomial term could be added to a linear regression model.
2. Generalized Linear Models (GLMs): If the relationship seems to be exponential, a GLM with a log link function could be appropriate.
3. Non-Parametric Regression: Methods like local regression (LOESS) could be used if the relationship is complex and not easily captured by standard parametric models.

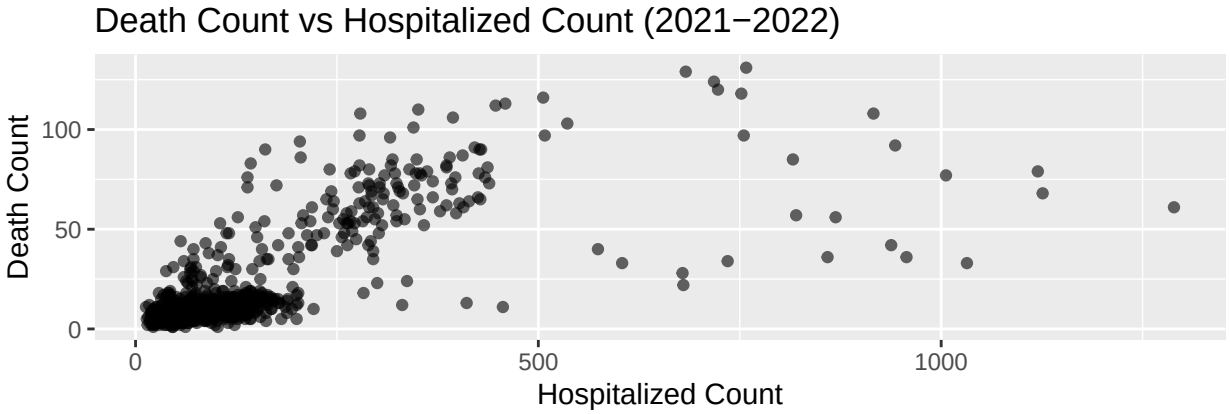
Data Appropriateness for Inference:

Relationship: The presence of a clear pattern in both plots is a good sign that there is a relationship to be modeled, but the non-linearity needs to be addressed. Variability: The spread of points around the 'line of best fit' will affect the model's predictive power. There seems to be increasing variability with higher counts, which suggests heteroscedasticity. Outliers: There do not appear to be as many extreme outliers in these scatter plots as in the histograms and box plots previously analyzed, but the dense clustering at the lower end could make it difficult to discern any outliers that might exist.

Independence: Assuming the data points (days) are independent of each other, this condition for inference is likely met, although if data are from time series, temporal autocorrelation should be checked. Before proceeding with model selection, it would be prudent to conduct further exploratory data analysis, including checking for overdispersion, examining residuals for any chosen model, and considering transformations or more complex models if necessary. For time-series data, accounting for temporal autocorrelation would be essential.

```
library(ggplot2)
library(gridExtra)

# Scatter plot for CASE_COUNT vs DEATH_COUNT
scatter_plot_cases <- ggplot(covid_data_2021_2022, aes(x=case_count, y=death_count)) +
  geom_point(alpha=0.6) +
  ggtitle('Death Count vs Case Count (2021-2022)') +
  xlab('Case Count') +
  ylab('Death Count')

# Scatter plot for HOSPITALIZED_COUNT vs DEATH_COUNT
scatter_plot_hospitalized <- ggplot(covid_data_2021_2022, aes(x=hospitalized_count, y=death_count)) +
  geom_point(alpha=0.6) +
  ggtitle('Death Count vs Hospitalized Count (2021-2022)') +
  xlab('Hospitalized Count') +
  ylab('Death Count')

# Arrange the plots side by side
grid.arrange(scatter_plot_cases, scatter_plot_hospitalized, ncol = 1)
```





Selection of Appropriate Regression Model Based on Visualizations:

Linear Regression Model Given the skewness of the data and the presence of outliers shown by the historams and box plots, a linear regression model might not be the best fit. The presence of outliers and the skewness indicate that the assumptions of normality and homoscedasticity (constant variance) for linear regression are likely violated. The Q-Q plots suggest that none of the variables follow a normal distribution, which is typical for count data like cases, hospitalizations, and deaths related to diseases. This non-normality means that assumptions required for certain statistical inference methods and for linear regression may not be met.

Poison Regression Model Given the overdispersion and skewness in the data, a Poisson regression model may not be appropriate because it assumes that the mean and variance of the data are equal. The overdispersion observed here can lead to underestimated standard errors and thus inflated test statistics, leading to incorrect inferences.

Negative Binomial Model A Negative Binomial regression model would likely be more appropriate for this data, as it can model the count data with overdispersion. It adds an extra parameter to account for the overdispersion, which can provide more reliable estimates of the standard errors and a better fit to the data.

Statistical Output

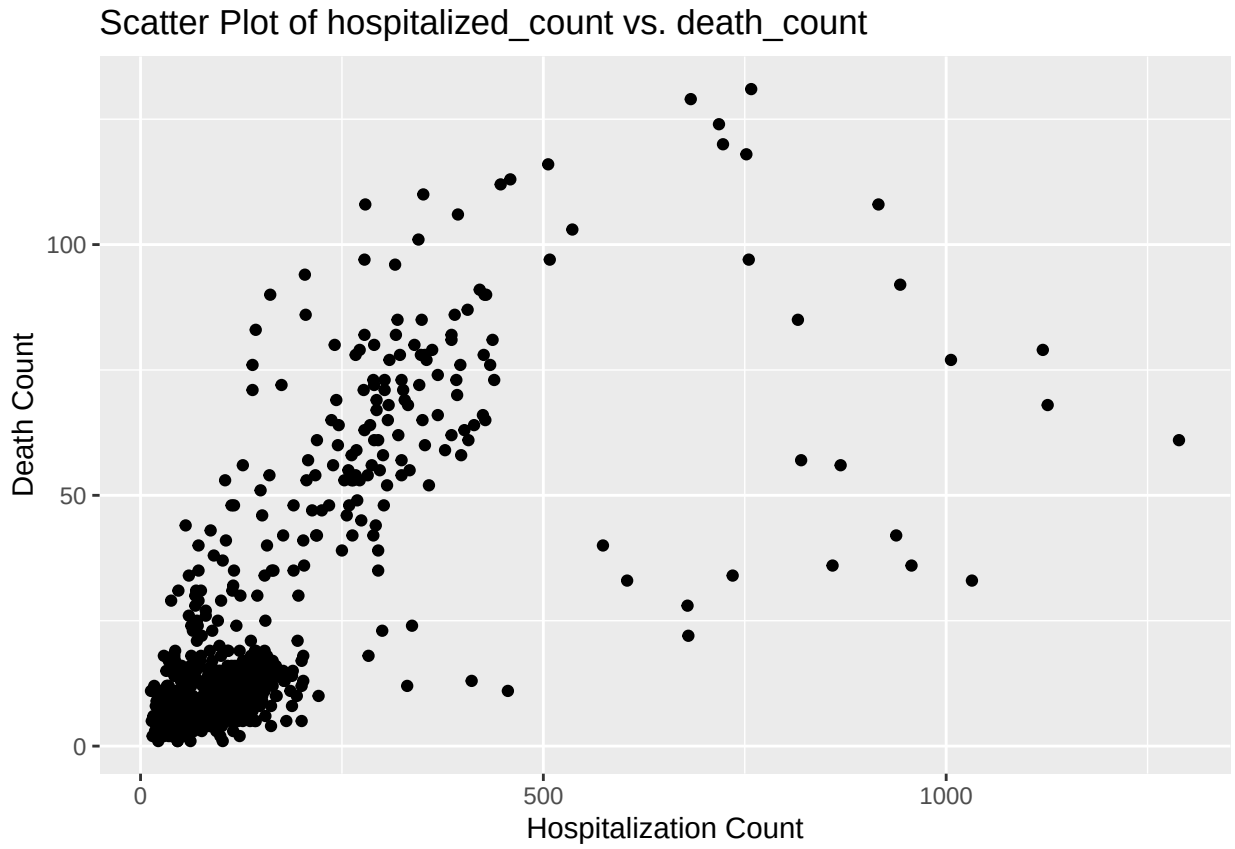
Hypothesis Test

Hypothesis Statement Null Hypothesis (H0): There is no statistical association between the number of hospitalizations ('hospitalized_count') and the number of deaths ('death_count') due to COVID-19. Any observed association in the data is due to random chance.

Alternative Hypothesis (H1): There is a statistically significant association between the number of hospitalizations ('hospitalized_count') and the number of deaths ('death_count') due to COVID-19. The observed association is not due to random chance.

Since the relationship appears to be linear in the scatter plot, I would be comfortable using a linear regression model to predict death_count based on hospitalized_count. To make the prediction, I would check for linearity by calculating the R^2, evaluating the residuals, and using statistical tests, to confirm the suitability of the linear model.

```
ggplot(covid_data_2021_2022, aes(x=hospitalized_count, y=death_count)) +
  geom_point() +
  labs(x = "Hospitalization Count", y = "Death Count") +
  ggtitle("Scatter Plot of hospitalized_count vs. death_count")
```



Since the relationship looks linear, we can quantify the strength of the relationship with the correlation coefficient.The correlation coefficient of 0.71 indicates that there is a strong relationship between death_count and hospitalized_count

Given a strong positive correlation coefficient of 0.71 and a very small p-value, we can conclude with high confidence that there is a significant and strong positive linear relationship between the number of hospitalizations and the number of deaths in the COVID-19 data. This result is statistically significant, meaning that it's highly unlikely to have occurred by chance.

Pearson's Correlation

```r
library(infer)
library(dplyr)

# Assuming covid_data_2021_2022 is your data frame
result <- covid_data_2021_2022 %>%
  summarise(
    correlation_coefficient = cor(hospitalized_count, death_count, use = "complete.obs"),
    p_value = cor.test(hospitalized_count, death_count, use = "complete.obs")$p.value
  )

# Displaying the results
# Output results
print(paste("Pearson's Correlation Coefficient:", result$correlation_coefficient))
```

```
## [1] "Pearson's Correlation Coefficient: 0.712425914462745"
```

```r
print(paste("P-value:", result$p_value))
```

```
## [1] "P-value: 2.18579769833151e-108"
```

```r
#print(result)
```

Spearman's Correlation

Since our data is skewed and have outliers, we will calculate the Spearman's correlation coefficient and p-value for comparison purpose. Spearman's correlation is a non-parametric measure of rank correlation, making it more appropriate for data that are not normally distributed or have outliers.

In comparison to the Pearsons correlation coefficient and p-value agrees with the Pearsons method ndicating that there is a strong relationship between death_count and hospitalized_count.

```r
suppressWarnings({
result <- cor.test(covid_data_2021_2022$death_count, covid_data_2021_2022$hospitalized_count, method =
})

# The correlation coefficient
correlation_coefficient <- result$estimate

# The p-value
p_value <- format(result$p.value, digits = 10, scientific = TRUE)

# Print the results
print(paste("Spearman's correlation coefficient:", correlation_coefficient))
```

```
## [1] "Spearman's correlation coefficient: 0.718000937845811"
```

```r
print(paste("P-value:", p_value))
```

```
## [1] "P-value: 7.699869412e-111"
```

Conclusion of Hypothesis Test:

Given the p-value is less than 0.05, we reject the Null Hypothesis (H0). There is statistically significant evidence to suggest an association between the number of hospitalizations and the number of deaths due to COVID-19. The strength and direction of this relationship, as indicated by the correlation coefficient suggest that as hospitalizations increase, deaths tend to increase as well, and this relationship is not due to random chance.

Important Considerations:

While we reject the null hypothesis and accept the alternative, it's crucial to remember that correlation does not imply causation. The observed association might be influenced by various other factors such as vaccination and public heath campaigns.

Regression

Simple Linear Regression Model

Next, we will fit a simple linear model to predict death count by hospitalization count.

```
# Applying a log-plus-one transformation to the variables
covid_data_2021_2022$log_death_count <- log1p(covid_data_2021_2022$death_count)
covid_data_2021_2022$log_hospitalized_count <- log1p(covid_data_2021_2022$hospitalized_count)
covid_data_2021_2022$log_case_count <- log1p(covid_data_2021_2022$case_count)

covid_death_simple <- lm(log_death_count ~ log_hospitalized_count, data = covid_data_2021_2022)
summary(covid_death_simple)
```

```
##
## Call:
## lm(formula = log_death_count ~ log_hospitalized_count, data = covid_data_2021_2022)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03794 -0.40617 -0.04724  0.49517  1.55200
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.00085    0.12753  -7.848 1.61e-14 ***
## log_hospitalized_count  0.80521    0.02688  29.961  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6124 on 691 degrees of freedom
## Multiple R-squared:  0.565,  Adjusted R-squared:  0.5644
## F-statistic: 897.6 on 1 and 691 DF,  p-value: < 2.2e-16
```

Equation of the simple linear regression

$$\widehat{death\_count} = -1.00085 + 0.80521 \times log\_hospitalized\_count$$

Interpretation of slope: For each one-unit increase in log_hospitalization count, the log_death_count is expected to increase by approximately 0.80521.

Statistical Significance: The p-value in the linear model is 2.2×10^-16. The p-value is significantly less than 0.05. This indicates that hospitalization count is a statistically significant predictor of the death count from Covid-19 in this dataset.

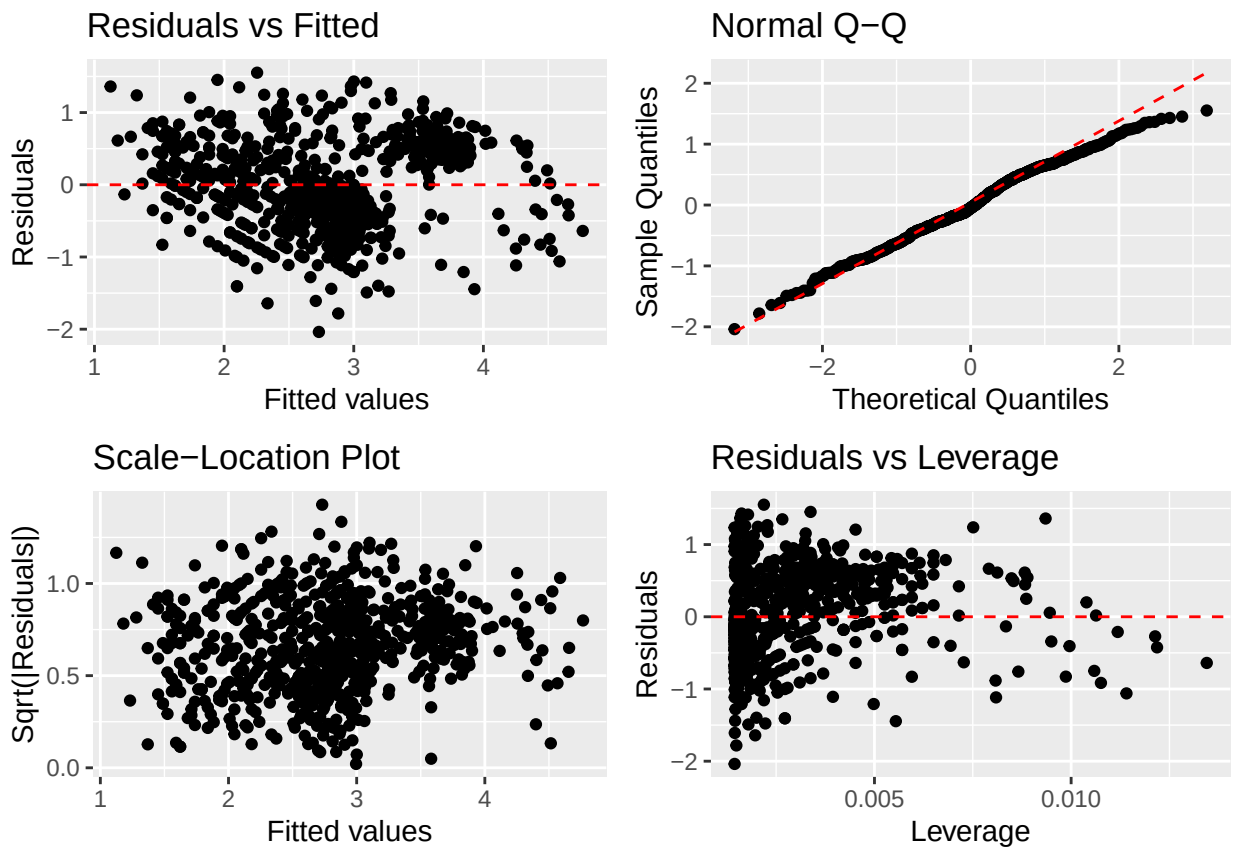
Practical Significance: 'Hospitalized_count' appears to be both a statistically and practically significant predictor of 'death_count' in the context of COVID-19 data. The model suggests a meaningful relationship between the number of hospitalizations and the number of deaths, highlighting the importance of hospitalization data in understanding and managing the impact of the COVID-19 pandemic.

```
# Add residuals and fitted values to the data frame
covid_data_2021_2022 <- covid_data_2021_2022 %>%
  mutate(fitted_values = fitted(covid_death_simple),
         residuals = resid(covid_death_simple),
         sqrt_abs_resid = sqrt(abs(residuals)),
         leverage = hatvalues(covid_death_simple))

# Creating the plots
p16_1 <- ggplot(covid_data_2021_2022, aes(x = fitted_values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype="dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted values", y = "Residuals")

p16_2 <- ggplot(covid_data_2021_2022, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(slope=1, intercept=0, color="red", linetype="dashed") +
  labs(title = "Normal Q-Q", x = "Theoretical Quantiles", y = "Sample Quantiles")
```

Simple Linear Model Diagnostics

```
## Warning in stat_qq_line(slope = 1, intercept = 0, color = "red", linetype =
## "dashed"): Ignoring unknown parameters: `slope` and `intercept`
```

```
p16_3 <- ggplot(covid_data_2021_2022, aes(x = fitted_values, y = sqrt_abs_resid)) +
  geom_point() +
  labs(title = "Scale-Location Plot", x = "Fitted values", y = "Sqrt(|Residuals|)")

p16_4 <- ggplot(covid_data_2021_2022, aes(x = leverage, y = residuals)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype="dashed") +
  labs(title = "Residuals vs Leverage", x = "Leverage", y = "Residuals")

grid.arrange(p16_1, p16_2, p16_3, p16_4, ncol = 2, nrow = 2)
```



Interpretation of diagnostic Plots of Simple Linear Regression

Residuals vs Fitted: The plot shows a funnel-like pattern where the spread of residuals increases with the fitted values, which is indicative of heteroscedasticity. This suggests that the variance of the errors is not constant and that a simple linear regression model may not be appropriate.

Normal Q-Q: The plot exhibits a slight deviation from the line at both tails, suggesting that the residuals may not be perfectly normally distributed. This could be indicative of outliers or skewness in the data.

Scale-Location (Spread-Location): The increasing spread from left to right indicates that the residuals have non-constant variance, which is consistent with the pattern observed in the Residuals vs Fitted plot.

Residuals vs Leverage: Most data points have low leverage, but there are a few with higher leverage. While these points have higher leverage, they do not appear to have large residuals, which suggests that they may not be unduly influencing the regression line.

Conclusion of diagnostic plots Given the above diagnostic results, a different type of regression model that accounts for heteroscedasticity and non-normality, might be more appropriate.

Multiple Linear Regression Model

```
# Applying a log-plus-one transformation to the variables
covid_data_2021_2022$log_death_count <- log1p(covid_data_2021_2022$death_count)
covid_data_2021_2022$log_hospitalized_count <- log1p(covid_data_2021_2022$hospitalized_count)
covid_data_2021_2022$log_case_count <- log1p(covid_data_2021_2022$case_count)

death_count_multiple <- lm(log_death_count ~ log_hospitalized_count + log_case_count, data = covid_data_
summary(death_count_multiple)
```

```
##
## Call:
## lm(formula = log_death_count ~ log_hospitalized_count + log_case_count,
##     data = covid_data_2021_2022)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58573 -0.34415 -0.00804  0.35975  1.43866
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.51164    0.14665   3.489 0.000516 ***
## log_hospitalized_count   1.42440    0.04603  30.942  < 2e-16 ***
## log_case_count          -0.59537    0.03826 -15.559  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5273 on 690 degrees of freedom
## Multiple R-squared:  0.678,  Adjusted R-squared:  0.6771
## F-statistic: 726.5 on 2 and 690 DF,  p-value: < 2.2e-16
```

Equation of the simple linear regression

$$\widehat{death\_count} = 0.51164 + 1.42440 \times hospitalized\_count - 0.59537 \times case\_count$$

Interpretation of Coefficients: 1. Hospitalized Count: The coefficient is 1.42440, indicating that for each additional hospitalization, the model predicts an increase of approximately 1.42440 deaths, holding 'case_count' constant. This is a sizable effect and suggests practical significance.

2. Case Count: The coefficient is -0.59537. This negative coefficient implies that for each additional case, the model predicts a slight decrease in the number of deaths, holding 'hospitalized_count' constant. This is counterintuitive and requires careful interpretation.

Statistical Significance: Both 'hospitalized_count' and 'case_count' have p-values much less than 0.05, indicating that their relationships with 'death_count' are statistically significant.

Practical Significance: The practical significance of 'hospitalized_count' is clear, given its positive and relatively large coefficient. The practical significance of 'case_count' is less clear due to the negative coefficient, which may suggest a more complex relationship. It might have been influenced by factors like healthcare capacity increase and treatment advancements, or varying severities of cases over time.

Model Fit: The 'Multiple R-squared' value is 0.678, meaning that about 6.78% of the variability in 'death_count' is explained by the model. This is a strong model fit, especially for observational data.

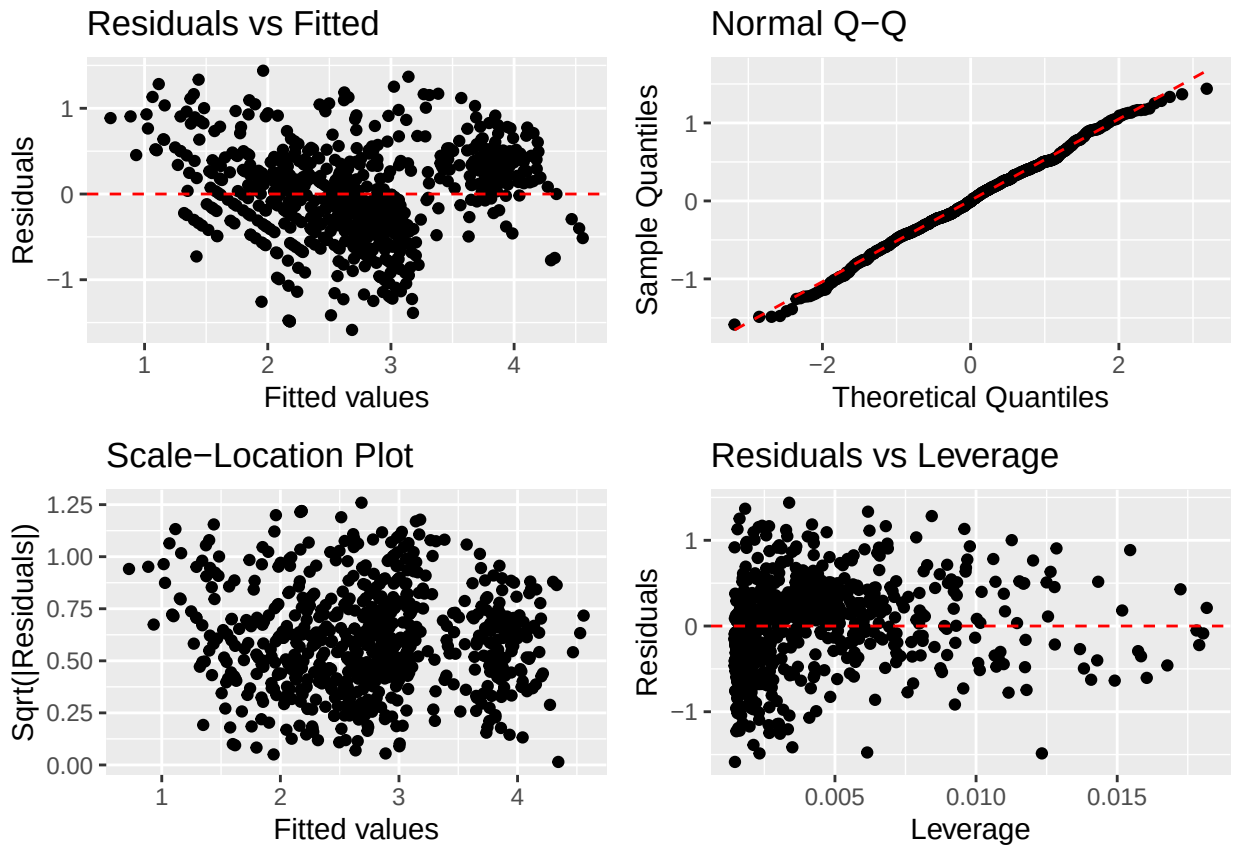
```r
# Add residuals and fitted values to the data frame
covid_data_2021_2022 <- covid_data_2021_2022 %>%
  mutate(fitted_values = fitted(death_count_multiple),
         residuals = resid(death_count_multiple),
         sqrt_abs_resid = sqrt(abs(residuals)),
         leverage = hatvalues(death_count_multiple))
```

```r
# Creating the plots
p16_1 <- ggplot(covid_data_2021_2022, aes(x = fitted_values, y = residuals)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype="dashed") +
  labs(title = "Residuals vs Fitted", x = "Fitted values", y = "Residuals")

p16_2 <- ggplot(covid_data_2021_2022, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(slope=1, intercept=0, color="red", linetype="dashed") +
  labs(title = "Normal Q-Q", x = "Theoretical Quantiles", y = "Sample Quantiles")
```

Multiple Linear Model Diagnostics

```
## Warning in stat_qq_line(slope = 1, intercept = 0, color = "red", linetype =
## "dashed"): Ignoring unknown parameters: ‘slope‘ and ‘intercept‘
```

```r
p16_3 <- ggplot(covid_data_2021_2022, aes(x = fitted_values, y = sqrt_abs_resid)) +
  geom_point() +
  labs(title = "Scale-Location Plot", x = "Fitted values", y = "Sqrt(|Residuals|)")

p16_4 <- ggplot(covid_data_2021_2022, aes(x = leverage, y = residuals)) +
  geom_point() +
  geom_hline(yintercept=0, color="red", linetype="dashed") +
  labs(title = "Residuals vs Leverage", x = "Leverage", y = "Residuals")

grid.arrange(p16_1, p16_2, p16_3, p16_4, ncol = 2, nrow = 2)
```



Interpretation of diagnostic Plots of Multiple Linear Regression Residuals vs Fitted plot The Residuals vs Fitted plot suggests possible deviations from linearity or the presence of heteroscedasticity, indicating that the model may not capture all the complexity of the data.

Normal Q-Q plot The Normal Q-Q plot indicates that the residuals are mostly normally distributed, which is a good sign for the reliability of coefficient estimates and the overall model.

Scale-Location plot The Scale-Location plot suggests that the variance of residuals is relatively constant, with potential slight heteroscedasticity for larger fitted values.

Residuals vs Leverage plot The Residuals vs Leverage plot does not show any particular points that would be considered highly influential or problematic for the model.

Overall Overall, while the diagnostic plots indicate the model fits the data reasonably well, there may be room for improvement, possibly through transforming variables, adding interaction terms, or considering non-linear models to better capture the underlying relationships in the data.

Generalized Linear Models (GLM)

Since the Linear Regression model does not seem to be appropiate for the count data, We will explore Generalized Linear Models. Generalized Linear Models (GLM) extend linear regression to models with a non-normal response distribution. For count data like 'death_count', a Poisson or negative binomial GLM can be appropriate since these distributions are commonly used for modeling count data.We will start with the Poison Model

Generalized Poisson model

We will use the GLM functions fit a GLM to the data with a Poisson error distribution, which is appropriate for count data.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
# Convert date_of_interest to Date type and filter for 2021 and 2022
covid_data$date_of_interest <- as.Date(covid_data$date_of_interest, format="%Y-%m-%d")
covid_data <- subset(covid_data, format(date_of_interest, "%Y") %in% c("2021", "2022"))

# Log-transform the independent variables
covid_data$log_hospitalized_count <- log1p(covid_data$hospitalized_count)
covid_data$log_case_count <- log1p(covid_data$case_count)

# Fit a Poisson GLM with a log link
poisson_model <- glm(death_count ~ log_hospitalized_count + log_case_count,
                     family = poisson(link = "log"), data = covid_data)

# Model summary
summary(poisson_model)
```

```
##
## Call:
## glm(formula = death_count ~ log_hospitalized_count + log_case_count,
##     family = poisson(link = "log"), data = covid_data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.11523    0.05871   1.963   0.0497 *
## log_hospitalized_count   1.72372    0.02066  83.431   <2e-16 ***
## log_case_count          -0.72909    0.01599 -45.592   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16114.3  on 692  degrees of freedom
## Residual deviance:  3731.8  on 690  degrees of freedom
## AIC: 6864.6
##
## Number of Fisher Scoring iterations: 5
```

Equation of the Poison Model

$$\log(\text{death\_count}) = 0.11523 + 1.72372 \cdot \log(\text{hospitalized\_count}) - 0.72909 \cdot \log(\text{case\_count})$$

To find the expected number of deaths (death_count), you would exponentiate both sides of this equation, leading to:

$$\widehat{\text{death\_count}} = \exp(0.11523 + 1.72372 \cdot \log(\text{hospitalized\_count}) - 0.72909 \cdot \log(\text{case\_count}))$$

Statistical Significance: The model shows statistical significance as indicated by the p-values., practical significance depends on the context.

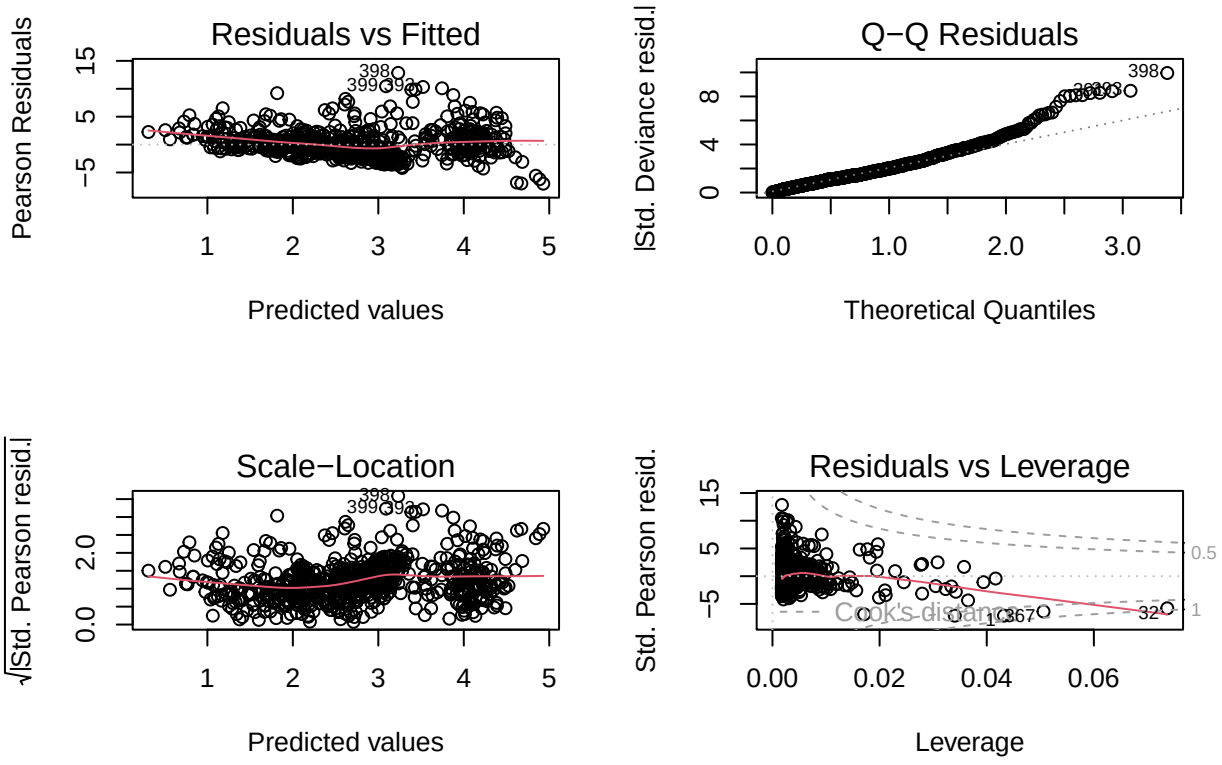
Practical Significance of the model: The coefficients for log_hospitalized_count and log_case_count are 1.72372 and -0.72909, respectively. These values are substantial, suggesting that changes in the predictors have a notable effect on the expected number of deaths. Particularly, the positive coefficient for log_hospitalized_count indicates a strong positive relationship with the expected death count. The negative coefficient for log_case_count is less intuitive but may suggest that as the overall number of cases increases, the proportion resulting in death decreases, possibly due to improved treatments, increased testing leading to the identification of milder cases.

Model Fit and Predictive Power: The drop from the null deviance to the residual deviance suggests a good model fit. However, the practical significance also depends on how well the model predicts new or unseen data. This would require validation using separate data.

Overall Interpretation While the model appears to have statistical significance and the coefficients suggest meaningful relationships, whether it is practically significant for predicting the expected number of deaths depends on the accuracy of its predictions in real-world scenarios and the context in which it is being used.

diagnostics for the Generalized Poisson model Below are the diagnostic plots. Below the plots are the interpretation of each plot.

```
# Diagnostic plots
par(mfrow = c(2, 2))
plot(poisson_model)
```



Residuals vs Fitted Plot: In this plot, there seems to be a funnel shape with a wider spread of residuals as the fitted values increase, indicating potential overdispersion or a non-linear relationship not captured by the model.

Normal Q-Q Plot: The points in the plot follow the line closely except for the upper tail (high quantiles), suggesting that high values may not be fitting as well, which could also be a sign of overdispersion.

Scale-Location Plot (or Spread-Location Plot): The plot indicates that variability of the residuals increases with the fitted values, another potential sign of overdispersion.

Residuals vs Leverage Plot: In this plot, there don't appear to be any points beyond the Cook's distance lines, suggesting there are no highly influential outliers.

Based on these diagnostics, there are signs of overdispersion in the data, which is common in count data. Overdispersion occurs when the variance is greater than the mean, which violates the Poisson assumption of equal mean and variance. This can lead to underestimating the standard errors of the coefficients, resulting in overly optimistic p-values.

Generalized Linear Model (GLM) with a negative binomial distribution

Next, we will use the Negative Binomial Generalized Linear Model (GLM) for the count data since there is evidence of overdispersion, which seems to be the case based on the above Poison model diagnostics. The

negative binomial model has an additional parameter to model the overdispersion, making it more flexible than the Poisson model for such data.

```r
library(MASS)

# Convert date_of_interest to Date type and filter for 2021 and 2022
covid_data$date_of_interest <- as.Date(covid_data$date_of_interest)
covid_data <- subset(covid_data, format(date_of_interest, "%Y") %in% c("2021", "2022"))

# Log-transform the independent variables (plus one to avoid log(0))
covid_data$log_hospitalized_count <- log1p(covid_data$hospitalized_count)
covid_data$log_case_count <- log1p(covid_data$case_count)

# Fit a Negative Binomial GLM with log-transformed independent variables
neg_binom_model <- glm.nb(death_count ~ log_hospitalized_count + log_case_count, data = covid_data)

# Model summary
summary(neg_binom_model)
```

```
##
## Call:
## glm.nb(formula = death_count ~ log_hospitalized_count + log_case_count,
##     data = covid_data, init.theta = 4.667059116, link = log)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.48015    0.14977   3.206  0.00135 **
## log_hospitalized_count   1.59602    0.04975  32.079  < 2e-16 ***
## log_case_count          -0.69407    0.04049 -17.143  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(4.6671) family taken to be 1)
##
##     Null deviance: 2627.13  on 692  degrees of freedom
## Residual deviance:  713.81  on 690  degrees of freedom
## AIC: 4896.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  4.667
##           Std. Err.:  0.326
##
##  2 x log-likelihood:  -4888.098
```

Equation of the GLM with Negative Binomial Disribution.

$$\widehat{\text{Death\_Count}} = \exp(0.48015 + 1.59602 \cdot \log(\text{Hospitalized\_Count}) - 0.69407 \cdot \log(\text{Case\_Count}))$$

Statistical Significance The low p-values (p-value = 0.00135) for the coefficients suggest that the model is statistically significant in predicting the death_count variable.

Practical Significance The magnitudes of the coefficients (1.59602 for hospitalized count and -0.69407 for case count) are substantial, especially given that they are on a logarithmic scale. This implies that changes in the log of hospitalized and case counts have meaningful impacts on the death count.

Model Fit 1. Null Deviance (2627.13): Represents the deviance of the model with only the intercept (no predictors). It's the baseline to compare. 2. Residual Deviance (713.81): Represents the deviance of the model with predictors. The significant drop from 2627.13 to 713.81 suggests a good fit. 3. AIC (Akaike Information Criterion) - 4896.1: Lower AIC values indicate a better model. The substantial decrease in deviance suggests a good fit.

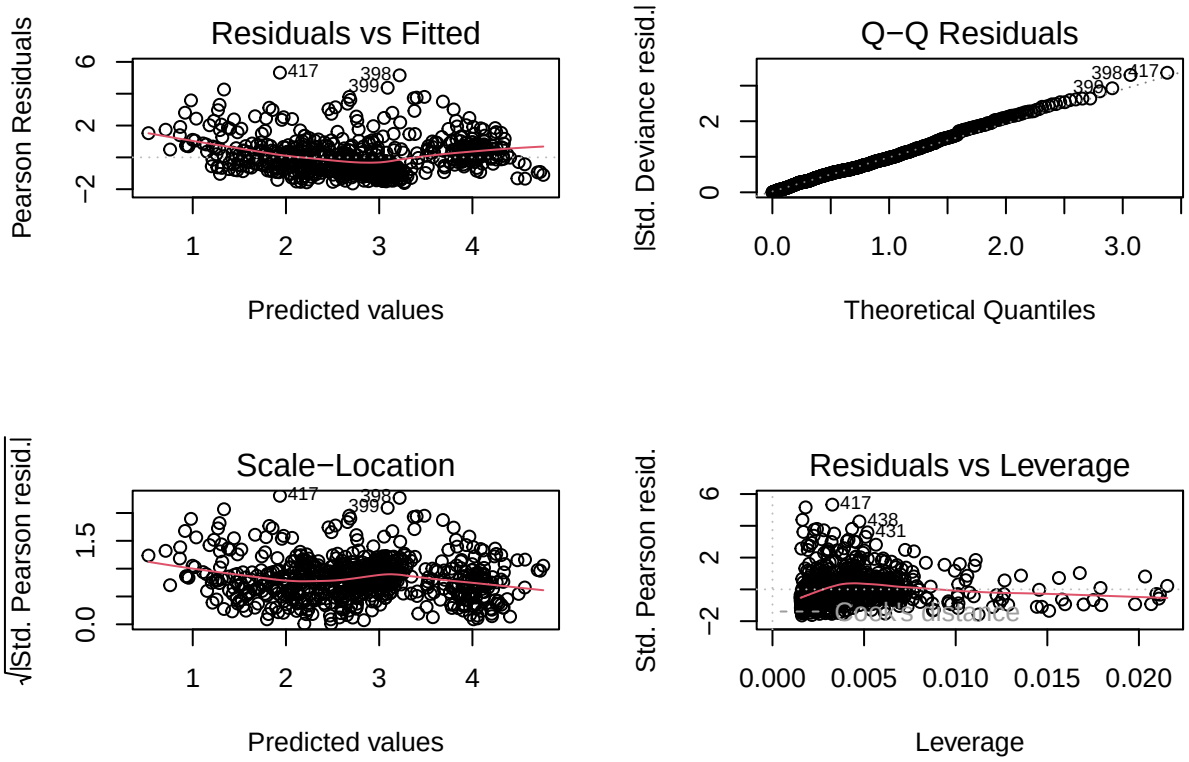
Predictive Power The low residual deviance relative to the null deviance and the significance of the coefficients suggest that the model has good predictive power.

Comparison with the Poison Model Theta (Dispersion parameter): A value of 4.667 with a standard error of 0.326 indicates that the negative binomial distribution is appropriate (suggesting overdispersion in the data which would not be well modeled by a Poisson distribution).

Overall Interpretation for GLM with Negative Binomial Distribution The model appears to be both statistically and practically significant in predicting death_count based on log_hospitalized_count and log_case_count. It fits the data well and should have good predictive power.

diagnostics for the Generalized Linear Model (GLM) with a negative binomial distribution Next we generate and interpret the diagnostic plots for the Generalized Linear Model (GLM) with a negative binomial distribution above. Below the plots is the interpretation each of the plots:

```
# Diagnostic plots
par(mfrow = c(2, 2))
plot(neg_binom_model)
```



Residuals vs Fitted: The plot shows a random scatter of residuals around the horizontal line, which is good. However, some points are lying outside the expected range, which could be outliers or indicate overdispersion.

Q-Q Residuals: The clear deviation at the upper tail indicates that the residuals have heavier tails than expected if they were normally distributed, which is common for count data and consistent with the negative binomial distribution's allowance for overdispersion.

Scale-Location: The plot indicates increasing spread with the fitted values, suggesting some degree of heteroscedasticity.

Residuals vs Leverage: The plot shows a few points with higher leverage, but they do not coincide with large residuals, suggesting they may not be having a significant influence on the model.

Overall Interpretation: The diagnostic plots indicate that while the negative binomial model may be accounting for overdispersion (as suggested by the residuals being better behaved than in the Poisson model), there are still some issues:

There might be outliers or points that are not well accounted for by the model, as indicated by the points lying outside the expected range in the Residuals vs Fitted plot and the heavy tails in the Q-Q plot.The variance of residuals might not be constant, as indicated by the Scale-Location plot. The Residuals vs Leverage plot does not raise immediate concerns about influential outliers, but it is always good practice to investigate any points with high leverage or high residuals to ensure they are not unduly affecting the model.

Conclusion

Importance of the analysis

This analysis is crucial as it enhances our understanding of COVID-19's impact and informs public health responses. By scrutinizing daily counts of cases, hospitalizations, and deaths from 2021 to 2022, the study aimed to identify a statistically robust model that accurately captures the relationships within the data. The Negative Binomial regression model emerged as the most appropriate, addressing the overdispersion and skewness typical in count data, which standard linear and Poisson models failed to accommodate.

Limitations of the analysis

However, the analysis has limitations. It did not account for variables like demographic factors or healthcare accessibility, which can influence COVID-19 outcomes. It was also limited to a specific timeframe and did not include time-series considerations, potentially overlooking trends and patterns over time. Furthermore, the assumption of independent observations may not hold true due to the interconnected nature of epidemiological data.

In summary, while the study improved the modeling of COVID-19 data by using a Negative Binomial regression, its insights are constrained by the scope of variables considered and the methodology's inherent assumptions. Future analysis could benefit from a more holistic approach, incorporating a broader range of factors and advanced modeling techniques.