Analyzing the Impact of Demographic, Socioeconomic, and Racial Factors on COVID-19 Risk and Outcomes in U.S. Counties

Group 4

Aakriti Bhandari, Sai Sathvik Appagana, Sai Srilekha Aluri, Sri Jahnavi Adusumilli

Department of Health Informatics, IUI

INFO B518 - Applied Statistical Methods for Biomedical Informatics

Dr. Gopikrishnan Chandrasekharan

December 16, 2024

Abstract:

The COVID-19 pandemic revealed profound health disparities across U.S. counties, driven by complex socioeconomic and demographic factors. This study aimed to investigate the relationships between poverty, race, demographic characteristics, and COVID-19 outcomes. Using the COVID-19 Race, Gender, and Poverty Risk Dataset covering 3,142 U.S. counties, we employed mixed statistical methods including exploratory data analysis, non-parametric tests, and regression modelling. Linear and logistic regression analyses were conducted to examine the impact of socioeconomic and demographic variables on COVID-19 deaths and health risk categorizations. Results demonstrated a significant positive correlation between poverty rates and COVID-19 mortality, with the linear regression model explaining 85% of death variability. The logistic regression model achieved 98.53% accuracy in predicting county health risk categories. Key findings highlighted disproportionate pandemic impacts, with counties experiencing higher poverty and larger minority populations suffering more severe outcomes. These results underscore the critical importance of socioeconomic factors in pandemic vulnerability and provide evidence-based insights for targeted public health interventions and equitable resource allocation strategies.

Keywords: COVID-19, health disparities, socioeconomic factors, regression analysis, public health

Introduction

The research focuses on understanding how socioeconomic, demographic, and racial factors influence COVID-19 outcomes across U.S. counties. It addresses two key questions:

1. **Research Question 1**: How do socioeconomic, racial, and demographic factors impact the cumulative number of COVID-19 deaths?
2. **Research Question 2**: How do these factors influence health risk categorizations across different counties and states?

Background and Significance

The COVID-19 pandemic has highlighted significant health disparities, with minority groups and those living in poverty being more severely affected (Hennis et al., 2021). These populations faced higher rates of infection, severe illness, and death due to unequal access to healthcare, poor living conditions, and limited preventive resources. Factors like poverty, population density, and racial diversity greatly influenced these disparities (Hennis et al., 2021). This study is important as it examines how these factors impact COVID-19 outcomes, offering insights to reduce health inequalities and guide public health policies. This study aims to identify ways to improve health equity and prepare for future crises.

Dataset and Scope

The analysis uses the COVID-19 Race, Gender, and Poverty Risk Dataset from Kaggle, which combines data from trusted sources like USA Facts, the U.S. Census, CDC, and Policy Map [1]. It contains 3,142 rows representing U.S. counties and 21 variables, covering factors like poverty rates, racial and gender demographics, health risk scores, and COVID-19 cases and deaths. The study explores how these factors impact health outcomes and risk levels. Using methods like Exploratory Data Analysis (EDA) to explore data trends, hypothesis testing to validate assumptions, regression analysis, and logistic regression, the goal is to uncover patterns that explain why some areas were hit harder by the pandemic. The findings aim to inform fair public health policies and prevention strategies.

Data Description

The dataset comes from Kaggle and combines reliable data from organizations like USA Facts, the U.S. Census Bureau, the CDC, and Policy Map [1]. It includes information on 3,142 U.S. counties and 21 variables that highlight socioeconomic, demographic, and health-related factors. These variables are crucial for analyzing the links between poverty, race, demographics, and COVID-19 outcomes. This dataset helps to explore health disparities between counties and can guide policies to address inequities and improve outcomes in future health crises.

Key Variables

1. **COVID-19 Deaths**: Cumulative deaths per 100,000 population in each county.
2. **COVID-19 Cases**: Total number of cases reported per county.
3. **Poverty Rate**: The proportion of the population living under the poverty threshold.
4. **Demographic Breakdown**: Includes variables such as White Male (W_Male), Black Female (B_Female), and others, reflecting population distributions by race and gender.
5. **Health Risk Index**: This represents the relative health risk in each county.
6. **Health Risk Category**: counties categories such as "Above Average" and "High Risk" etc.
7. **County and State**: Geographic identifiers for each record.

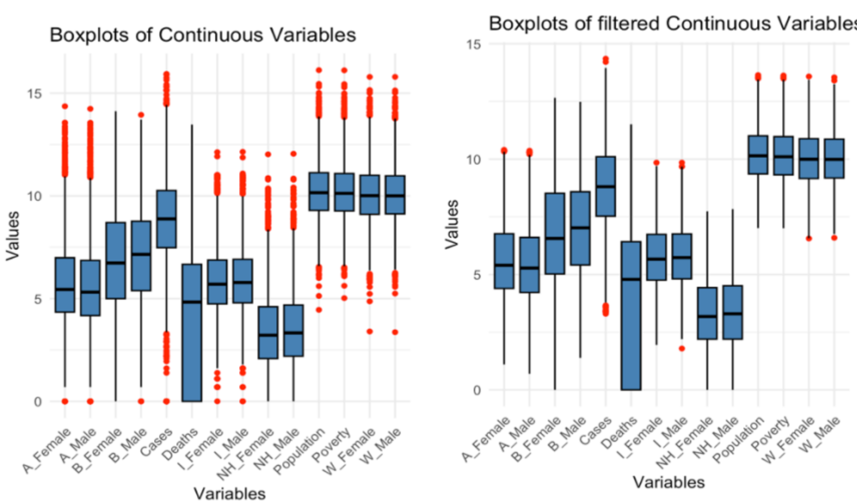
Table 1 *Classification of Variables*

	CATEGORICAL		QUANTITATIVE	
	NOMINAL	ORDINAL	DISCRETE	CONTINUOUS
INDEPENDENT VARIABLES	County, State	-	-	Poverty rate, Demographic breakdown of Race and Gender, Health Risk Index, COVID-19 Cases
DEPENDENT VARIABLES	-	Health Risk Category	-	COVID-19 Deaths

Preprocessing Steps: These steps ensure the dataset is clean, structured, and ready for analysis.

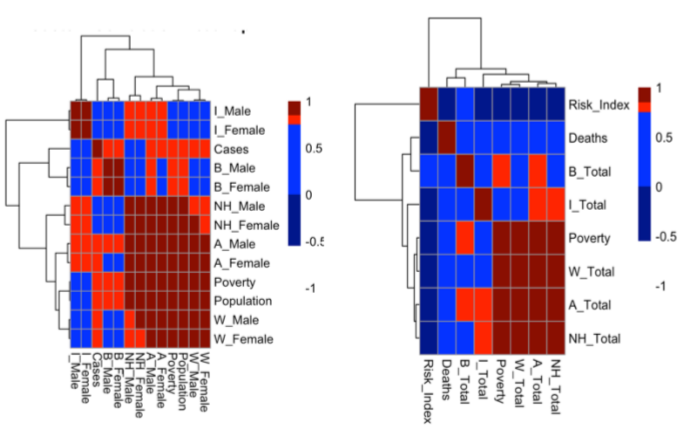
1. **Missing Data**: The dataset was evaluated for null values using the colSums(is.na(data)) function, confirming that no missing data was present.
2. **Irrelevant Columns**: Three initial columns unrelated to the research question were dropped to streamline the dataset and focus on variables that contributed to the analysis.
3. **Removal of Outliers**: Outliers are removed using the IQR method, creating a cleaner representation of the core data by focusing on the main dataset's variability.

Figure 1 *Box plots before and after removing outliers*



4. **Multicollinearity**: Correlations between highly related variables (male and female population totals) were addressed by aggregating these groups and removing redundant columns (Population).

Figure 2 *Correlation heatmap of the variables*



Descriptive Statistics and Visualizations

Descriptive Statistics: Using the summary () function, descriptive statistics are measured for the numerical variables. Central tendency measures (mean and median) revealed typical values for deaths, cases, and poverty rates. Variability (minimum, maximum, interquartile range) highlighted significant disparities between counties, particularly for deaths and cases.

Figure 3 *Descriptive statistics of the numerical variables*

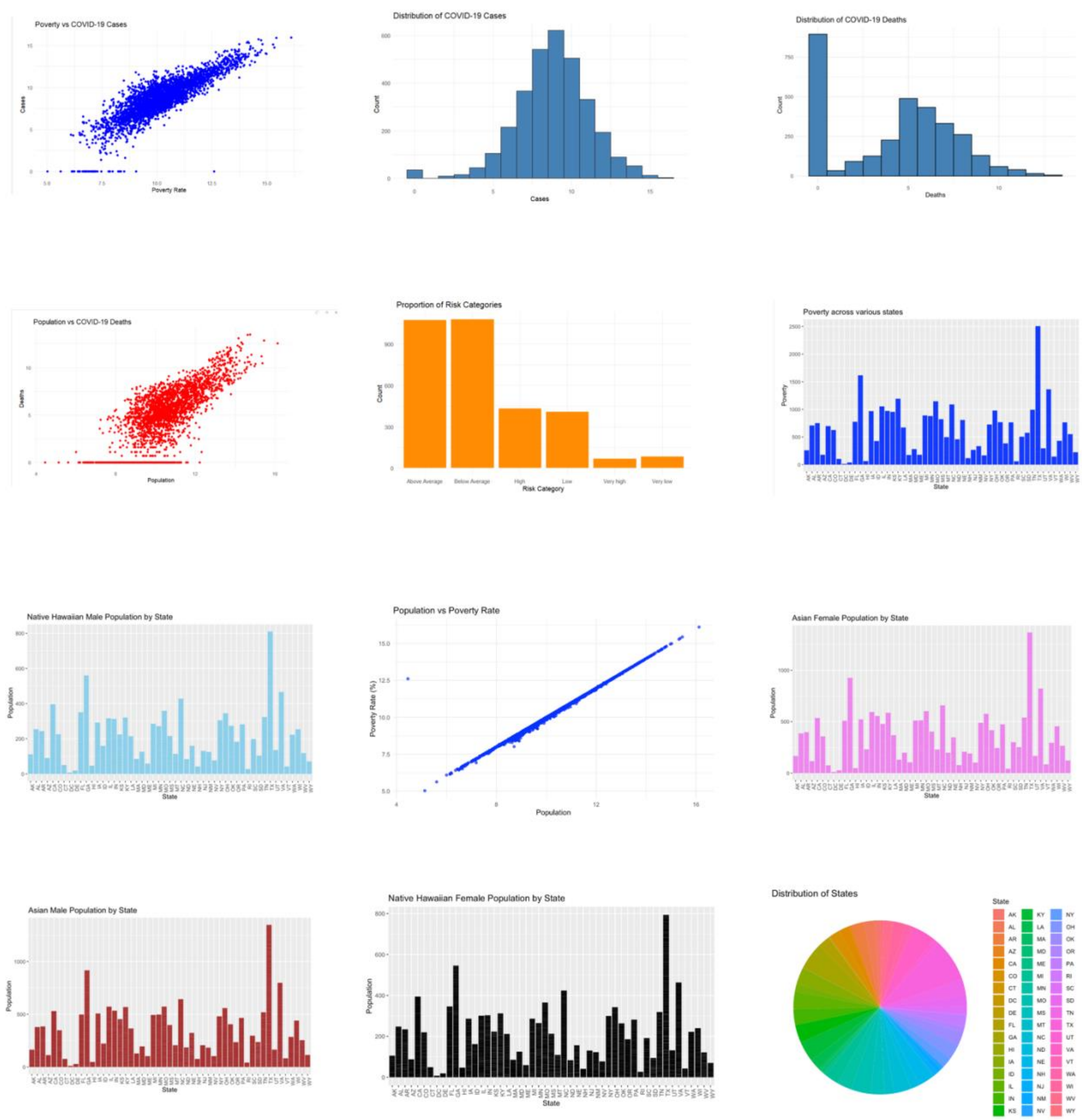
```
     Cases            Deaths           Poverty          Population          W_Male          W_Female
 Min.   : 0.000   Min.   : 0.000   Min.   : 5.024   Min.   : 4.454   Min.   : 3.367   Min.   : 3.401
 1st Qu.: 7.478   1st Qu.: 0.000   1st Qu.: 9.269   1st Qu.: 9.297   1st Qu.: 9.125   1st Qu.: 9.104
 Median : 8.880   Median : 4.832   Median :10.124   Median :10.155   Median :10.003   Median :10.012
 Mean   : 8.814   Mean   : 4.269   Mean   :10.241   Mean   :10.274   Mean   :10.089   Mean   :10.084
 3rd Qu.:10.263   3rd Qu.: 6.671   3rd Qu.:11.090   3rd Qu.:11.128   3rd Qu.:10.982   3rd Qu.:11.005
 Max.   :15.939   Max.   :13.473   Max.   :16.113   Max.   :16.122   Max.   :15.795   Max.   :15.794
     B_Male          B_Female          I_Male           I_Female          A_Male           A_Female
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 5.389   1st Qu.: 4.999   1st Qu.: 4.796   1st Qu.: 4.745   1st Qu.: 4.174   1st Qu.: 4.344
 Median : 7.153   Median : 6.732   Median : 5.784   Median : 5.704   Median : 5.318   Median : 5.447
 Mean   : 7.116   Mean   : 6.876   Mean   : 5.933   Mean   : 5.887   Mean   : 5.649   Mean   : 5.812
 3rd Qu.: 8.768   3rd Qu.: 8.702   3rd Qu.: 6.916   3rd Qu.: 6.879   3rd Qu.: 6.859   3rd Qu.: 6.987
 Max.   :13.947   Max.   :14.125   Max.   :12.146   Max.   :12.130   Max.   :14.241   Max.   :14.365
    NH_Male          NH_Female        Risk_Index      Poverty_column   Population_column
 Min.   : 0.000   Min.   : 0.000   Min.   :  2.85   Min.   : 5.024   Min.   : 4.454
 1st Qu.: 2.197   1st Qu.: 2.079   1st Qu.: 27.13   1st Qu.: 9.269   1st Qu.: 9.297
 Median : 3.332   Median : 3.219   Median : 51.38   Median :10.124   Median :10.155
 Mean   : 3.474   Mean   : 3.415   Mean   : 51.32   Mean   :10.241   Mean   :10.274
 3rd Qu.: 4.689   3rd Qu.: 4.605   3rd Qu.: 75.47   3rd Qu.:11.090   3rd Qu.:11.128
 Max.   :12.050   Max.   :12.024   Max.   :100.00   Max.   :16.113   Max.   :16.122
```

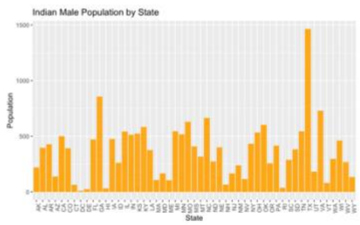
Key Findings from Visualizations:

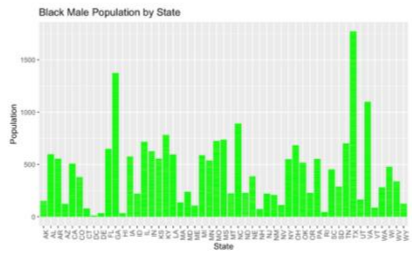
1. **Distribution of COVID-19 Cases and Deaths**: Histograms showed that most counties had few to moderate cases and deaths, while some areas had much higher numbers, indicating the pandemic affected regions differently. Deaths were uneven, with most counties having few or no deaths, but some had more.
2. **COVID-19 Cases and Poverty Rates**: Scatter plots showed a strong link between higher poverty rates and more COVID-19 cases, indicating poorer areas were more impacted.
3. **Risk Categories Across Counties**: Bar charts revealed that most counties were categorized as "Above Average" or "Below Average Risk," with fewer as "High Risk" or "Very High Risk" aligning with disparities in healthcare access and socioeconomic factors.

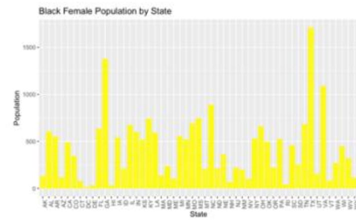
4. **Population and Poverty Across States**: The scatter plot demonstrated a strong linear relationship between population size and poverty rates. States with larger populations, such as Texas, showed higher poverty levels. The pie chart visualized the distribution across states.
5. **Demographic Distribution**: Bar charts displayed the racial breakdown across states, showing that counties with larger Black and Hispanic populations had significantly higher COVID-19 cases & deaths.

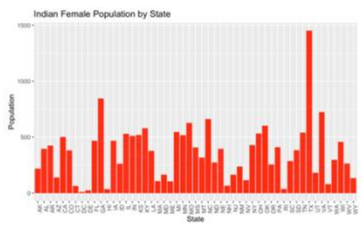
Figures 4, 5, and 6 *Visualizations of Individual Variables and Their Relationships*

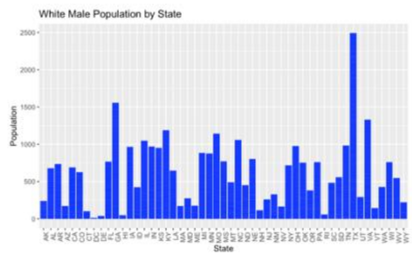


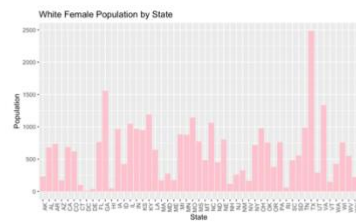












Statistical Methods

Normality testing: The Shapiro-Wilk test checked data normality, and a p-value < 0.05 indicated non-normality. This is supported using non-parametric tests, which handle skewed data and outliers well.

Non-Parametric Tests: The Kruskal-Walli's test was used to compare numeric variables (e.g., poverty, COVID-19 cases, deaths) across risk categories. Significant differences ($p < 0.05$) highlighted disparities in health outcomes between counties. The Chi-square test examined relationships between categorical variables (risk categories and regions), showing strong associations and geographic disparities in COVID-19 impact.

Correlation Analysis (Spearman's Rank Correlation): Spearman's test identified strong positive correlations between poverty and COVID-19 outcomes and between population size and deaths.

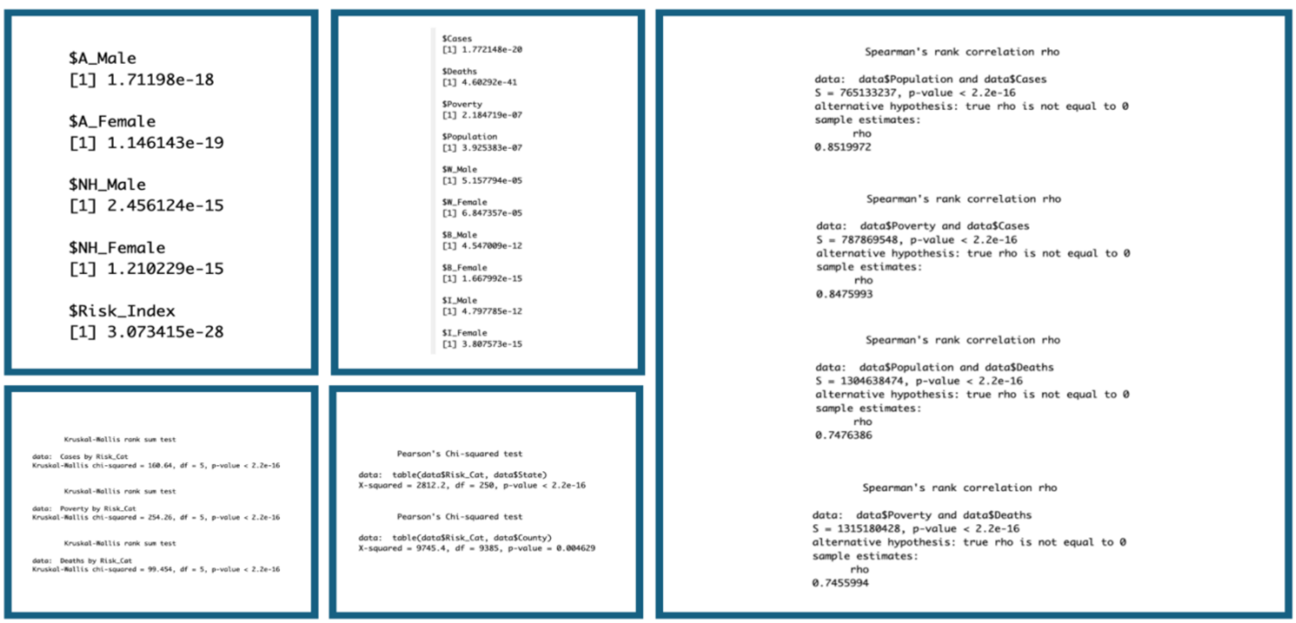
Regression Models (Linear and Logistic): Linear regression was employed to predict COVID-19 deaths using independent variables like poverty, cases, health risk index, and racial demographics. Logistic regression was used for effectively predicting health risk categories (e.g., "Above Average," "High Risk") based on demographic and socioeconomic factors.

Justification and Relevance: The statistical methods were chosen based on the data and research goals. The Shapiro-Wilk test confirmed that non-parametric methods like Kruskal-Wallis and Spearman correlation were suitable for non-normal and skewed data. These methods helped compare variables and analyze relationships. Regression models examined how predictors affected outcomes, offering insights into the factors behind COVID-19 deaths and health risks. Together, these methods provided clear and accurate results to guide public health strategies.

Results

The Kruskal-Wallis tests showed differences in Cases, Poverty, and Deaths across risk categories. Chi-square tests found strong links between risk categories and geography, like State and County. Spearman's Rank Correlation found that poverty and population size were linked to more COVID-19 cases and deaths, showing how socioeconomic and demographic factors affected the pandemic.

Figure 7 *Normality Assessment, Statistical Analysis, and Model Results*



```
$A_Male
[1] 1.71198e-18

$A_Female
[1] 1.146143e-19

$NH_Male
[1] 2.456124e-15

$NH_Female
[1] 1.210229e-15

$Risk_Index
[1] 3.073415e-28
```

```
$Cases
[1] 1.772148e-20

$Deaths
[1] 4.60292e-41

$Poverty
[1] 2.184719e-07

$Population
[1] 3.925383e-07

$W_Male
[1] 5.157794e-05

$W_Female
[1] 6.847357e-05

$B_Male
[1] 4.547009e-12

$B_Female
[1] 1.667992e-15

$I_Male
[1] 4.797785e-12

$I_Female
[1] 3.807573e-15
```

```
        Kruskal-Wallis rank sum test

data:  Cases by Risk_Cat
Kruskal-Wallis chi-squared = 160.64, df = 5, p-value < 2.2e-16

        Kruskal-Wallis rank sum test

data:  Poverty by Risk_Cat
Kruskal-Wallis chi-squared = 254.26, df = 5, p-value < 2.2e-16

        Kruskal-Wallis rank sum test

data:  Deaths by Risk_Cat
Kruskal-Wallis chi-squared = 99.454, df = 5, p-value < 2.2e-16
```

```
        Pearson's Chi-squared test

data:  table(data$Risk_Cat, data$State)
X-squared = 2812.2, df = 250, p-value < 2.2e-16

        Pearson's Chi-squared test

data:  table(data$Risk_Cat, data$County)
X-squared = 9745.4, df = 9385, p-value = 0.004629
```

```
        Spearman's rank correlation rho

data:  data$Population and data$Cases
S = 765133237, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8519972

        Spearman's rank correlation rho

data:  data$Poverty and data$Cases
S = 787869548, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8475993

        Spearman's rank correlation rho

data:  data$Population and data$Deaths
S = 1304638474, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7476386

        Spearman's rank correlation rho

data:  data$Poverty and data$Deaths
S = 1315180428, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7455994
```

The linear regression model explained 85% of the variation in deaths (adjusted R square of 0.8475), with significant predictors being poverty (positively related), cases (positively related), and the health risk index (negatively related). Diagnostics like residual plots and Cook's Distance confirmed the model was reliable. The logistic regression model had an accuracy of 98.53% and a Confidence Interval of (0.9713 to 0.9936), with high sensitivity and specificity, effectively predicting risk categories based on demographics and socioeconomic factors. Both models showed how key factors impact health and risk.

Diagnostic Plots: The Residuals vs. Fitted Plot. Residuals are randomly distributed around zero, but a slight funnel shape suggests potential heteroscedasticity. The Normal Q-Q Plot. Most residuals align with the diagonal line, but slight deviations at the tails indicate potential outliers. The Scale-Location Plot. The downward slope of the red line suggests heteroscedasticity, as residual variance decreases with fitted values. Cook’s Distance Plot. Larger bubble sizes highlight potential influential points.

Figure 8 *Diagnostics Plots suggesting heteroscedasticity*

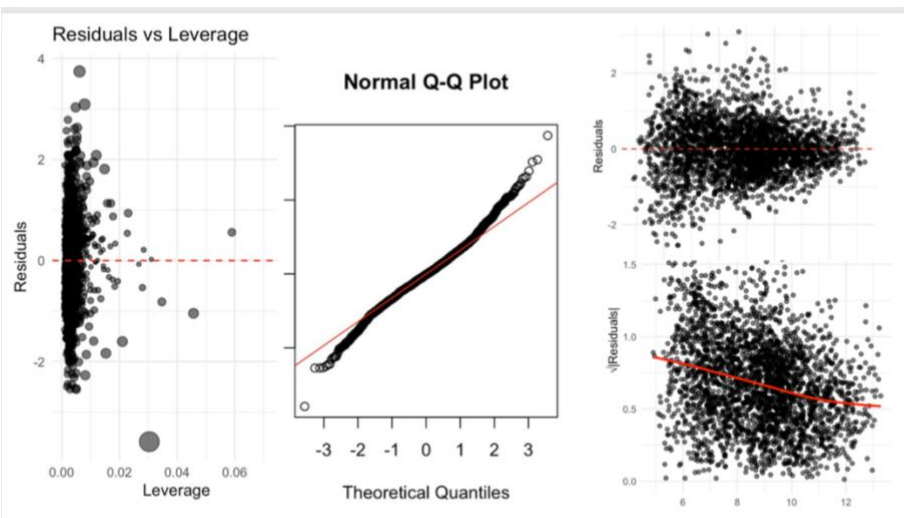


Figure 9 *Results of Linear and Logistic Regression*

```
Call:
lm(formula = Deaths ~ Poverty + Cases + Risk_Index + W_Total +
    B_Total + A_Total + I_Total + NH_Total, data = Aggregated_filtered_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5851 -0.4380 -0.0161  0.4310  3.7416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.7975240  0.2612120   6.881 7.32e-12 ***
Poverty      0.5426860  0.0994216   5.458 5.24e-08 ***
Cases        0.2342747  0.0071636  32.703  < 2e-16 ***
Risk_Index  -0.0022783  0.0006959  -3.274 0.001074 **
W_Total     -0.0571564  0.0383629  -1.490 0.136370
B_Total      0.1326813  0.0082242  16.133  < 2e-16 ***
A_Total     -0.0091113  0.0145242  -0.627 0.530501
I_Total     -0.0310380  0.0092061  -3.371 0.000758 ***
NH_Total     0.0607666  0.0109713   5.539 3.34e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7373 on 2731 degrees of freedom
Multiple R-squared:  0.8475,	Adjusted R-squared:  0.847
F-statistic:  1897 on 8 and 2731 DF,  p-value: < 2.2e-16
```

```
Prediction      Above Average Below Average High Low Very high Very low
  Above Average           197             1    3   0         0        0
  Below Average             0           184    0   0         0        0
  High                      0             0   79   0         1        0
  Low                       0             2    0  57         0        1
  Very high                 0             0    0   0        11        0
  Very low                  0             0    0   0         0        9

Overall Statistics

               Accuracy : 0.9853
                 95% CI : (0.9713, 0.9936)
    No Information Rate : 0.3615
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9795

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Above Average Class: Below Average Class: High Class: Low Class: Very high Class: Very low
Sensitivity                        1.0000               0.9840      0.9634     1.0000          0.91667         0.90000
Specificity                        0.9885               1.0000      0.9978     0.9939          1.00000         1.00000
Pos Pred Value                     0.9801               1.0000      0.9875     0.9500          1.00000         1.00000
Neg Pred Value                     1.0000               0.9917      0.9935     1.0000          0.99813         0.99813
Prevalence                         0.3615               0.3431      0.1505     0.1046          0.02202         0.01835
Detection Rate                     0.3615               0.3376      0.1450     0.1046          0.02018         0.01651
Detection Prevalence               0.3688               0.3376      0.1468     0.1101          0.02018         0.01651
Balanced Accuracy                  0.9943               0.9920      0.9806     0.9969          0.95833         0.95000
>
> # Calculate accuracy
> accuracy <- confusion_matrix$overall['Accuracy']
> cat("Accuracy of the Classification Model:", accuracy, "\n")
Accuracy of the Classification Model: 0.9853211
>
```

Discussion

The linear regression results show that poverty is positively linked to COVID-19 deaths, highlighting the greater impact on disadvantaged communities, while the negative relationship with the Risk Index suggests that improved risk management reduces fatalities. The model's strong explanatory power is reflected in the adjusted R^2 of 0.8475, though slight heteroscedasticity requires attention. In logistic regression, the model achieved 98.53% accuracy in predicting COVID-19 risk, with confidence intervals confirming reliability. Balanced accuracy shows the model's effectiveness in classifying categories, even with imbalanced data.

Significance of our Findings: The findings reveal that socio-economic and racial predictors significantly influence COVID-19 deaths and risk categorization. These insights can inform public health interventions and prioritize resource allocation for high-risk communities.

Unexpected Outcomes: The unexpected heteroscedasticity seen in the Scale-Location Plot could affect standard errors. Future models may need robust standard errors or variable changes.

Limitations of the Study

Removal of Outliers: Eliminating outliers might reduce variability and overlook extreme cases.

Correlation, Not Causation: The study shows correlations, not causality. For example, poverty is linked to higher COVID-19 cases and deaths, but it doesn't prove a direct cause-and-effect relationship.

Data Distribution Assumptions: While non-parametric tests are robust for non-normal data, they are less sensitive than parametric tests, potentially limiting the detection of small but meaningful differences.

Conclusion: The analysis highlighted significant COVID-19 disparities across U.S. counties, with socioeconomic and racial factors impacting outcomes. The linear regression model found a link between poverty and mortality, explaining 85% of death variability, while the logistic regression model achieved 98.53% accuracy. These findings highlight the need for targeted interventions, fair resource distribution, and addressing social health factors to reduce future pandemic risks.

References

Covid 19 Race Gender Poverty Risk (U.S County). *Kaggle.* Retrieved from Covid 19 Race Gender Poverty Risk (U.S County)

Hennis, A. J. M., Coates, A., Del Pino, S., Ghidinelli, M., Gomez Ponce de Leon, R., Bolastig, E., Castellanos, L., Oliveira E Souza, R., & Luciani, S. (2021). COVID-19 and inequities in the Americas: Lessons learned and implications for essential health services. *Pan American Journal of Public Health*, *45*, e130. https://doi.org/10.26633/RPSP.2021.130