Multiple Linear Regression Project: Socio-Economic Factors Influencing The Cancer Mortality

Aleksandra Kalisz

Postgraduate Diploma in Science in Data Analytics
National College of Ireland
Dublin, Ireland
21118876@student.ncirl.ie

***Abstract* —** This report investigates the socio-economic factors influencing cancer mortality in the US. The purpose of this research is to identify the most significant predictors of cancer mortality using a multiple linear regression model. The findings can be used to prioritize cancer prevention, used for early detection, and affect and potentially reduced mortality rates overall.

Introduction

In this report, we will use multiple linear regression to understand the relationship between a dependent variable and multiple independent variables. This statistical method identifies the independent variable which has the strongest relationship with the dependent variable and uses them to build a predictive model[i]. The formula for this method is:

$$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c.$$

Figure 1. Multiple Regression Equation

Where 'y' is the dependant variable and 'x1', 'x2', … are independent variables, and 'b1', 'b2, … are coefficients that represent the effects of each independent variable on the dependent variable.
To build a successful model it is important to examine all data. For this purpose, we will be using R Studio which provides easy-to-use graphical interface and visualization tools. We will use descriptive statistics and visualizations to better understand variables and based on this knowledge we will be able to choose our predictors. We will check for collinearity between variables using a correlation matrix and heatmap. Based on our observations we will build our linear model, removing the outliers which can lead to inaccurate estimates of parameters.

The Assumptions

Multiple linear regression is a technique that relies on several assumptions to ensure the validity and reliability of results. Our assumptions are the following[ii] :

- There is a linear relationship between a dependent variable and independent variables (Linearity).
- The observations in the dataset are independent of each other (Independence).
- The residuals in the model are normally distributed (Normality).
- The variance of the errors is constant across all levels of the independent variables (homoscedasticity).
- The independent variables are not highly correlated with each other(No multicollinearity).

To address these assumptions, we will perform diagnostics like scatter plots, Durbin Watson test, Cook's distance, Normality test, variance inflation factor (VIF), and Residuals vs Fitted Plot. By performing these diagnostics, we can ensure that the assumptions of the model are met and that the multiple regression model is valid and reliable. We will take the necessary steps to address any assumptions that are proven to be incorrect. Finally, we will apply our model to calculate the "Mortality Rate" which will be a summary of division variables: "death rate" and "incidence rate".

In the conclusion of our findings, we will summarize the key socioeconomic factors that contribute to the death rate in the United States. We will discuss the relationships between these factors and the death rate highlighting the significance of addressing socioeconomic differences in healthcare. We'll also talk about our study's weaknesses and potential directions for further investigation.

Exploratory Data Analysis

Data scientists use exploratory data analysis (EDA) to analyse and investigate data sets and summarize their key characteristics, often using data visualization techniques. It aids in determining how to best manipulate data sources to obtain the answers required, making it simpler for data scientists to find patterns, detect anomalies, test hypotheses, and validate assumptions[iii].

The Data

For our analysis, we were provided with data on cancer mortality in the US for 3000+ counties. Data is presented in a .csv file and consist of 3047 rows and 25 columns. The variables in this data set are County and its population, death rate, incidence rate, median income, poverty percent, median age, median age male, median age female, average household size, percentage of married households, percentage of no high school age group 18-24,
percentage of the group with bachelor's degree 18 – 24, percentage of the group with a high school over 25, percentage of the group with bachelor's degree 25 and over, percentage of group unemployed 16 and over, percentage of the group with private coverage, percentage of the group with private coverage by employee percentage of the group with public coverage, percentage of the group with public coverage only, as well as the race-related variables like White, Black, Asian and other race.

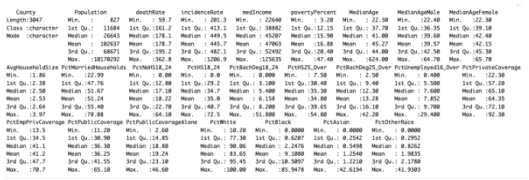

Figure 2. Summary Data

Data Exploration

Exploring data, we have noticed that some of the County's population in our data Population is smallest than 1000. The smallest population in our data set is 827 in Golden Valley County and the largest is 10170292 in Los Angeles County. As the incidence rate and death rate diagnostics are calculated per 100000 we decided to eliminate smaller populated counties by filtering the Population by larger than 10000. Filtering data can improve the accuracy and validity of the analysis. Also, to avoid bias in the data or inaccurate results, data were checked for missing values, but no missing values were found.

```
cancer_data <- as.data.frame(read.csv("cancer.csv", header = TRUE))
# sorting the data frame by the Population variable in ascending order
cancer_data_sorted <- cancer_data[order(cancer_data$Population),]
# displaying the sorted data frame
head(cancer_data_sorted)
tail(cancer_data_sorted)
```

	County	Population		County	Population
	<chr>	<int>		<chr>	<int>
64	Golden Valley County, Montana	827	2391	Orange County, California	3169776
414	Esmeralda County, Nevada	829	2672	San Diego County, California	3299521
116	Wibaux County, Montana	1130	2611	Maricopa County, Arizona	4167947
89	Sheridan County, North Dakota	1310	2851	Harris County, Texas	4538028
188	Greeley County, Kansas	1330	2570	Cook County, Illinois	5238216
700	Issaquena County, Mississippi	1337	2610	Los Angeles County, California	10170292

Figure 3. Sorted Data

```
data <- filter(cancer_data_sorted, Population > 10000 )
```

Figure 4. Data Filter

Visualisations and Correlations

Histograms of the dependent variable "deathRate" and each independent variable were created to identify potential issues with the normality of the distribution of the dependent variable and to identify outliers. Some of the variables had highly skewed distribution with a long tail on the left or right side. To address this, the proper transformations of the variables were made to improve the linearity and normality of the distribution. Using transformed variables for the model will improve its fit ensuring that the assumptions are met.

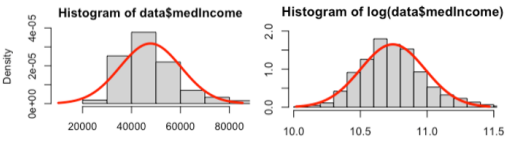

Figure 5. Histograms

Scatterplot Matrix and Heatmap

A scatterplot matrix was created to study the relationship between different variables. This function can help to identify any patterns or dependencies between independent variables. In one of the off-diagonal plots, we are searching for a linear relationship between the independent and dependent.

Considering the number of variables and the difficulty of reading small scatterplots we decided to create a Heatmap which makes it easy to visually identify the strength and direction of the relationship between variables.

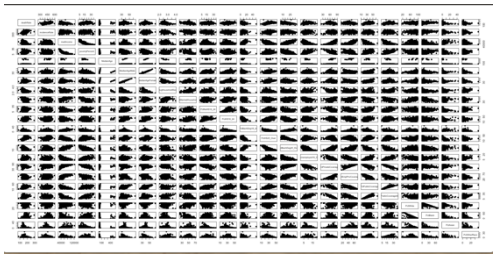

Figure 6. Scatterplot Matrix

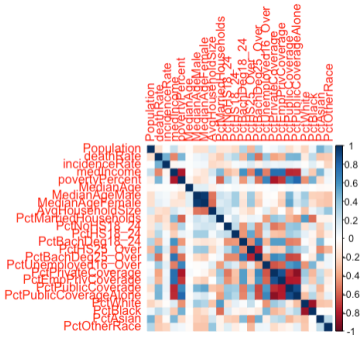

Figure 7. Multicollinearity Heatmap

We are looking for positive correlations marked as darker shades of blue, which indicates that variables are related, and might be useful predictors in our model. Also, we are looking for a strong negative correlation which is marked with red color and that indicates that variables are inversely related, and they also can be useful predictors. Most of all we are looking for the variables strongly correlated with multiple other variables as these variables might be affected by multicollinearity which can cause problems with the regression model. Squares light colored or white are not useful in our model. From our heatmap, we can see that the "death rate" variable is strongly correlated with "incidence rate" as "PctHS25Over","PovertPercent", "PctUnemployed16Over","PctPublicCoverAlone", "PublicCover", "PctBlack" and "medIncome".

The Model

The first step is to build a multiple linear model with all the independent variables. In this function, we are looking at Residuals if they are roughly centred around zero and if they are in a similar spread on either side. Also, we are looking at the regression coefficient of the model: The *Estimate* tells us about the strength and direction of the relationship between the predictor variable and response variable. *Std.Error* displays the error of the estimate, t- *value* shows test statistics. In other words, the larger the number is than less likely is that the results occurred by chance. The amount of proof contradicting the null hypothesis is indicated by the p-value.

```
Residuals:
    Min      1Q  Median      3Q     Max
-84.830 -10.780   0.438  10.597 155.385

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.682e+02  1.927e+01  13.924  < 2e-16 ***
povertyPercent          3.389e-01  1.721e-01   1.969 0.049044 *
medIncome               4.227e-04  8.982e-05   4.706 2.67e-06 ***
MedianAge              -7.046e-03  9.072e-03  -0.777 0.437415
MedianAgeMale          -8.740e-02  2.897e-01  -0.302 0.762903
MedianAgeFemale        -7.470e-01  3.065e-01  -2.437 0.014880 *
AvgHouseholdSize       -3.098e+01  3.505e+00  -8.838  < 2e-16 ***
PctMarriedHouseholds    5.486e-02  1.197e-01   0.458 0.646815
PctNoHS18_24           -1.799e-01  7.343e-02  -2.450 0.014348 *
PctBachDeg18_24         1.755e-01  1.459e-01   1.203 0.228980
PctHS25_Over            5.650e-01  1.087e-01   5.199 2.18e-07 ***
PctBachDeg25_Over      -1.835e+00  1.811e-01 -10.129  < 2e-16 ***
PctUnemployed16_Over    1.172e+00  1.985e-01   5.907 3.98e-09 ***
PctPrivateCoverage     -2.548e-01  1.705e-01  -1.494 0.135222
PctEmpPrivCoverage      1.597e-01  1.215e-01   1.315 0.188741
PctPublicCoverage      -4.653e-01  2.757e-01  -1.687 0.091643 .
PctPublicCoverageAlone  1.327e+00  3.481e-01   3.812 0.000141 ***
PctWhite               -3.493e-02  7.546e-02  -0.463 0.643466
PctBlack                1.471e-01  7.294e-02   2.017 0.043774 *
PctAsian               -6.986e-02  1.997e-01  -0.350 0.726453
PctOtherRace           -1.530e+00  1.534e-01  -9.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.8 on 2384 degrees of freedom
Multiple R-squared:  0.4627,    Adjusted R-squared:  0.4582
F-statistic: 102.6 on 20 and 2384 DF,  p-value: < 2.2e-16
```

Figure 9. All Variables Linear Model

As an example, in this model, we can take variable PctHS25_Over which indicates that for a one-unit increase in the percentage of individuals in a county who are over 25 years old and have completed high school, we would expect to see, on average, a 0.565 unit increase in the response variable, holding all other variables constant.

In this linear model, we must consider the value of Adjusted R2, which takes into account the number of predictors in multiple linear regression. The adjusted R2 value ranges from 0 to 1, with higher values showing that the model fits the data better.
In this case, the R2 value is quite small, and we will be looking to improve it by selecting only variables with significant values for our next model.
To improve our next model, we will add the transformed variables from earlier stages of visualizations to our dataset.

```
data$logIncome <- log(data$medIncome)
data$logPctBachDeg25_Over<- log(data$PctBachDeg25_Over)
data$sqrtPctUnemployed16_Over <- sqrt(data$PctUnemployed16_Over)
data$sqrtPctOtherRace <- sqrt(data$PctOtherRace)
```

```
 [1] "County"              "Population"          "deathRate"           "incidenceRate"       "medIncome"
 [6] "povertyPercent"      "MedianAge"           "MedianAgeMale"       "MedianAgeFemale"     "AvgHouseholdSize"
[11] "PctMarriedHouseholds" "PctNoHS18_24"        "PctHS18_24"          "PctBachDeg18_24"     "PctHS25_Over"
[16] "PctBachDeg25_Over"   "PctUnemployed16_Over" "PctPrivateCoverage"  "PctEmpPrivCoverage"  "PctPublicCoverage"
[21] "PctPublicCoverageAlone" "PctWhite"         "PctBlack"            "PctAsian"            "PctOtherRace"
[26] "logIncome"           "logPctBachDeg25_Over" "sqrtPctUnemployed16_Over" "sqrtPctOtherRace"
```

Figure 10. Variable Transformation

Based on observation and analysis, we have decided to include only the variables with a statistically significant value in our multiple regression model. This approach allowed us to eliminate any unnecessary or irrelevant variables and focus only on the factors that have a significant impact on the response variable. By doing so, we aimed to improve the accuracy and reliability of our model's predictions and ensure that the model is not overfitted.
Our chosen variables are:
incidenceRate, MedianAgeFemale, AvgHouseholdSize, logPctBachDeg25_Over, sqrtPctUnemployed16_Over, PctPublicCoverageAlone, PctWhite, sqrtPctOtherRace

```
Call:
lm(formula = deathRate ~ incidenceRate + MedianAgeFemale + AvgHouseholdSize +
    logPctBachDeg25_Over + sqrtPctUnemployed16_Over + PctPublicCoverageAlone +
    PctWhite + sqrtPctOtherRace, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-113.873   -9.439    0.059    9.371  131.834

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             230.607196  10.739508  21.473  < 2e-16 ***
incidenceRate             0.184137   0.006903  26.676  < 2e-16 ***
MedianAgeFemale          -0.676969   0.090336  -7.494 9.35e-14 ***
AvgHouseholdSize        -14.949705   2.028686  -7.369 2.35e-13 ***
logPctBachDeg25_Over    -26.636381   1.218489 -21.860  < 2e-16 ***
sqrtPctUnemployed16_Over  2.911837   0.939493   3.099  0.00196 **
PctPublicCoverageAlone    0.588891   0.096748   6.087 1.34e-09 ***
PctWhite                 -0.160645   0.026950  -5.961 2.88e-09 ***
sqrtPctOtherRace         -5.584644   0.495677 -11.267  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.71 on 2396 degrees of freedom
Multiple R-squared:  0.5738,    Adjusted R-squared:  0.5724
F-statistic: 403.2 on 8 and 2396 DF,  p-value: < 2.2e-16
```

Figure 8. New Observation Linear Model

From the summary above we can see that the Adjusted R2 has improved, and multiple R squared shows that the model explains 57.38% of the variation of the variance in the response variable. A p-value less than 0.05 indicates that the coefficient is statistically significant and there is evidence of a relationship between that predictor variable and the response variable.

By plotting our model, we can generate 4 separate plots:

- Residuals vs Fitted - plot aids in determining whether there is any pattern in the residuals that indicates a violation of the assumptions of linearity, constant variance, and error normality. In general, we want the plot to show no pattern and for the residuals to be randomly distributed around the zero line.
- Normal Q-Q diagram - to verify the residual's normal distribution. If the residuals are normally distributed the plot is linear.
- Scale- Location used to check if there is any pattern in the residuals' spread. We want the points to be randomly scattered around the horizontal line, suggesting constant variance of the residuals.
- The residuals vs. leverage plot help to identify influential observations that could have a significant effect on the regression results. We want the points to be evenly distributed and not clustered in any specific region of the plot.

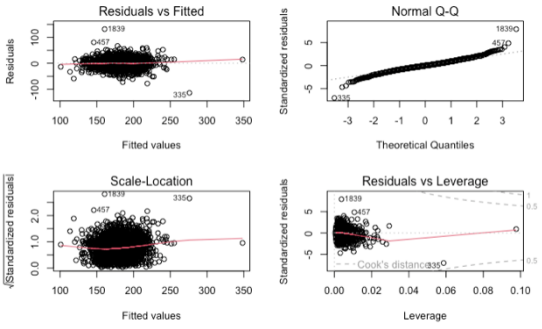

Figure 11. Linear Model Plot

In this plot, we can identify outliers that fall far away from horizontal zero which indicates that model has difficulty predicting the variance of the observations. The next step is to remove outliers and improve our model.

```
# Removing outliers
data1 <- subset( data, subset= ! as.integer(rownames(data)) %in% c (  549, 109, 951, 643, 109, 2310, 1091, 79, 1634, 338, 1187, 969,
940, 685, 585, 940, 276, 2320, 509, 352, 481, 1693, 212, 335, 1839, 1663, 498, 130, 285, 2363, 286, 457, 344, 1839, 1528, 615, 2026,
107,268, 305, 1665, 1666, 373, 1862, 2080, 238, 1384, 906, 1894, 1699, 1668, 54, 1598))
```

Figure 12. Outliers

After removing outliers, we can summarise our model again and see if the R squared has improved.

```
Call:
lm(formula = deathRate ~ incidenceRate + MedianAgeFemale + AvgHouseholdSize +
    logPctBachDeg25_Over + sqrtPctUnemployed16_Over + PctPublicCoverageAlone +
    PctWhite + sqrtPctOtherRace, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-61.282  -9.367  -0.206   9.151  58.649

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             221.11719   10.35912  21.345  < 2e-16 ***
incidenceRate             0.19292    0.00723  26.683  < 2e-16 ***
MedianAgeFemale          -0.71335    0.08621  -8.275  < 2e-16 ***
AvgHouseholdSize        -11.65445    2.02683  -5.750 1.01e-08 ***
logPctBachDeg25_Over    -26.59225    1.16824 -22.763  < 2e-16 ***
sqrtPctUnemployed16_Over  3.06162    0.90414   3.386  0.00072 ***
PctPublicCoverageAlone    0.61474    0.09305   6.607 4.85e-11 ***
PctWhite                 -0.18062    0.02654  -6.804 1.29e-11 ***
sqrtPctOtherRace         -6.07633    0.48454 -12.540  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.71 on 2346 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.6011
F-statistic: 444.4 on 8 and 2346 DF,  p-value: < 2.2e-16
```

Figure 13. Improved Linear Model

The above model demonstrates that there is a substantial relationship between the death rate and the independent variables. The coefficient estimates for each variable show how much the death rate varies as each independent variable changes while the other variables remain constant. Plotting our improved model below we can see more outliers that we could remove and improve our Rsquared.

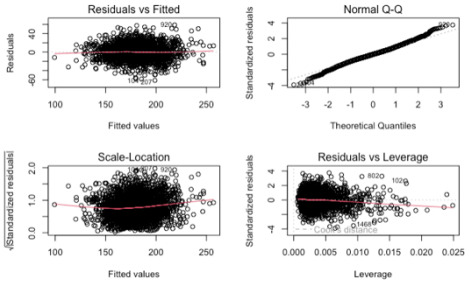

Figure 14. Improver Linear Model Plots

In the first plot of our improved model Residuals vs Fitted appears to be no noticeable trend in the residuals in this case, indicating that the model is doing a good job of fitting the data.

The second Normal Q-Q plot - the residuals, in this case, follow a straight line, yet there are minor departures from normality near the tails.

Third Scale – The location plot shows reveals that the residuals lack any clear pattern which suggests that the model has a constant variance.

In the last plot Residuals vs Leverage, there are some points outside the dashed lines that may have a substantial influence on model fit and may be outliers.

Cooks Distance Test

This test is a measure of the influence of each observation on the regression coefficients and can be used to identify potential outliers or influential observations.

	StudRes	Hat	CookD
207	-3.9254616	0.006480898	0.011100400
802	3.2414233	0.013783217	0.016651592
920	3.7539308	0.005535536	0.008667295
937	0.7477742	0.023946214	0.001524555
1468	-3.1366987	0.012592646	0.013889626
2162	-0.7586833	0.024873857	0.001631697

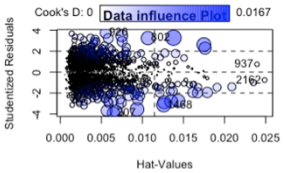

Figure 15. Cooks Distance Test

Looking at the Cooks distance values we can see that there are observations with particularly high values, indicating that it may be an influential data point and may still have some influence on the model. It is important to investigate these observations further to determine whether they are truly influential and whether they should be removed from the analysis.

Based on the data overall, the linear regression model appears to have a good fit, with an adjusted R-squared value of 0.6011 and a substantial F-statistic with a very low p-value. With p-values of 0.05, the coefficients for all predictors are also significant. However, before taking any conclusions, it is critical to thoroughly evaluate the model assumptions which we mentioned in the introduction.

Model evaluation and checking assumptions.

Linearity Test – Linear Plot

Checking for linearity when evaluating a model assumption ensures that the relationship between the variables is accurately represented, which leads to better model performance and more accurate predictions.

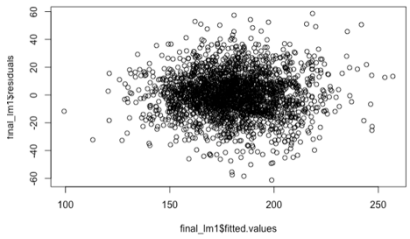

Figure 16. Linear Plot

In our plot, the fact that residuals are mostly concentrated around 0 indicates that the model is predicting the dependent variable reasonably well and that the predicted values are close to the actual values. It appears that the linearity assumption is likely being met.

Independence Test - Durbin Watson Test

Durbin Watson test is a test for autocorrelation in statistical models or regression analysis residuals.

This statistic has a constant value between 0 and 4. A value of 2.0 implies that no autocorrelation was discovered in the data. Values ranging from 0 to less than 2 indicate positive autocorrelation, whereas values ranging from 2 to 4 indicate negative autocorrelation[iv].

```
        Durbin-Watson test

data:  final_lm1
DW = 1.9716, p-value = 0.2413
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 17. Durbin Watson Test

The test resulted in a DW statistic of 1.9716 and a p-value of 0.2413. Establishing that the null hypothesis for the test is that there is no autocorrelation in the residuals. As the alternative hypothesis is that there is positive autocorrelation.

Since the p-value is greater than the significance level of 0.05, we fail to reject the null hypothesis, meaning that there is no evidence of positive autocorrelation in the residuals. Therefore, we can assume that the residuals are independent and do not violate the assumption of no autocorrelation.

Normality Test – Normal QQ Plot

In this test, the residuals are plotted against what would be predicted if they were regularly distributed.

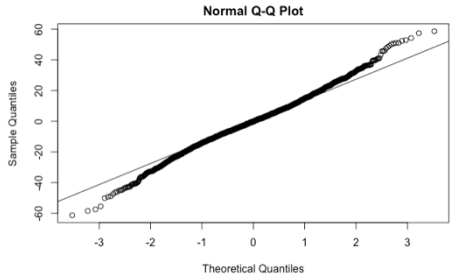

Figure 18. Normal QQ Plot

Our plot shows that there are slight deviations from normality close to the tails and this can be an indication that certain outliers or extremely high numbers are the sources of the variance.

Homoscedasticity Test – Plot of Residuals vs Fitted

This is an important assumption for linear regression models because if the assumption is violated, the standard errors and confidence intervals for the regression coefficients may be biased, leading to an inaccurate conclusion.

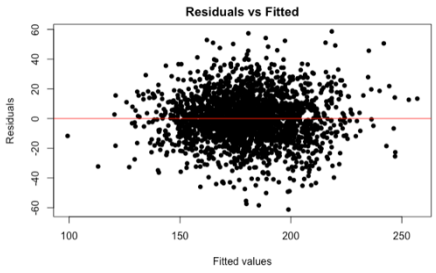

Figure 19. Residuals vs Fitted Plot

Our plot indicates homoscedasticity, which means that the variance of the residuals is relatively constant across all levels of the independent variables.

Multicollinearity occurs when predictor variables in a linear regression model are highly correlated with each other. AVIF value of 1 indicates no multicollinearity, while values above 1.

No multicollinearity Test - Variance Inflation Factor Test

The VIF test measures how the relationship between predictor variables affects the reliability of the estimated regression coefficients for each variable.

```
incidenceRate          MedianAgeFemale       AvgHouseholdSize     logPctBachDeg25_Over sqrtPctUnemployed16_Over   PctPublicCoverageAlone
     1.122760                 1.709480                1.812938                2.103865                2.029395                2.711644
      PctWhite          sqrtPctOtherRace
     1.543030                 1.493015
```

Figure 20. VIF Test

Our results show that there is some degree of multicollinearity among the independent variables, as all of the VIF values are greater than 1 but there are no values above 5 which would indicate that the predictor is highly correlated in the model. In general, the VIF test advises taking caution when interpreting the findings of a multiple regression analysis using these variables.

Homoscedasticity Test – Residuals vs Fitted Plot

A homoscedasticity test checks if the variability of the residuals is constant across the predictor variable(s) in a regression model. Violation of this assumption can affect the accuracy of the coefficients and the results.

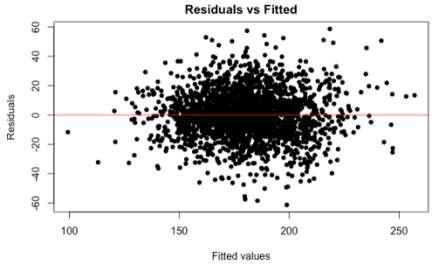

Figure 21. Homoscedasticity Test

Our results show a random scatter of points with no clear pattern, which indicates that the model is capturing the relationship between the variables well and that the residuals are normally distributed with a constant variance.

Model Summary

Based on the Residuals vs Fitted figure, which exhibits no obvious trend in the residuals, the modified model appears to match the data well. The residuals are normally distributed, as per the Normal Q-Q plot, with a few small outliers at the tails. According to the Scale-Location plot, the model shows constant variance. The Residuals vs Leverage plot, however, suggests that there might be some significant outliers in the data that might have an impact on the model's fit. Our outliers could be in the Population variable but that needs to be investigated further. Although the model seems to be a good match overall, further

research may be required to resolve any potential outliers or residual deviations from the norm.

The assumption of linearity appears to be met since the residuals in the plot are mostly concentrated around 0, indicating that the model is predicting the dependent variable reasonably well and that the predicted values are close to the actual values. The residuals do not show any signs of positive autocorrelation, according to the Durbin-Watson test's non-significant p-value, showing that they are independent and satisfy the assumption of no autocorrelation. According to the Q-Q plot, there are small deviations from normality at the tails, which might be an indication of outliers or exceptionally high values that contribute to variance. Overall, the residuals' variance appears to be rather stable, indicating that the homoscedasticity criterion has been met.

Mortality Rate Model

We decided to apply our model to “mortality rate” which was calculated by dividing “death rate” by” incidence rate” using the same predictors.

```
# calculating  mortality rate by dividing death rate by incidence rate
mortalityRate <- data1$deathRate / data1$incidenceRate

Call:
lm(formula = mortalityRate ~ incidenceRate + MedianAgeFemale +
    AvgHouseholdSize + logPctBachDeg25_Over + sqrtPctUnemployed16_Over +
    PctPublicCoverageAlone + PctWhite + sqrtPctOtherRace, data = data1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.148793 -0.021038  0.000133  0.020093  0.195933

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.146e-01  2.404e-02  38.051  < 2e-16 ***
incidenceRate           -4.978e-04  1.678e-05 -29.673  < 2e-16 ***
MedianAgeFemale         -1.466e-03  2.000e-04  -7.331 3.14e-13 ***
AvgHouseholdSize        -2.692e-02  4.703e-03  -5.725 1.17e-08 ***
logPctBachDeg25_Over    -6.060e-02  2.711e-03 -22.357  < 2e-16 ***
sqrtPctUnemployed16_Over  7.214e-03  2.098e-03   3.439 0.000594 ***
PctPublicCoverageAlone   1.272e-03  2.159e-04   5.893 4.35e-09 ***
PctWhite                -4.234e-04  6.159e-05  -6.875 7.94e-12 ***
sqrtPctOtherRace        -1.391e-02  1.124e-03 -12.375  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03645 on 2346 degrees of freedom
Multiple R-squared:  0.5207,    Adjusted R-squared:  0.5191
F-statistic: 318.6 on 8 and 2346 DF,  p-value: < 2.2e-16
```

Figure 22. Mortality Rate Formula and Model

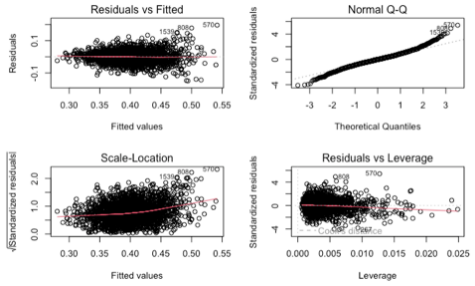

Figure 23. Mortality Rate Model Plots

This model behaves like our previous model. Coefficients show the direction and degree of independent variables linked to the dependent variable, and all have a strong correlation with the dependent variable based on the p-value. The R-squared of 0.5207 explains how much the independent variables can account for the variation in the dependent variable, whereas the residual standard error of 0.037 measures the average distance of data points from the regression line. Overall, the model is significant, and the independent variables have a strong collective relationship with the dependent variable.

Conclusion

In this project, we used a multiple regression model to identify the most influential socioeconomic factors that contribute to the death rate in the United States. Our analysis revealed eight factors that had a significant impact on the death rate, including incidence rate, median female age, average household size, percentage of the population over age 25 with a bachelor's degree or higher, percentage of the population unemployed age 16 and over, percentage of the population with public health coverage only, and percentage of the white race with the percentage of other races.

Most of these factors were related to poverty, education, and insurance coverage. From our findings, we can conclude that lack of education and low economic status are significant contributors to the death rate in the US. However, it's important to note that this is a broad conclusion. As there are some limitations in our dataset, further research is needed to fully understand the factors that contribute to the death rate in the US. Future studies could explore additional variables such as access to healthcare, social support networks, and environmental factors to gain a more comprehensive understanding of the underlying factors that affect mortality rates in the US population.

Overall, our analysis provides valuable insights into the factors that contribute to the death rate in the US and highlights the importance of addressing socioeconomic gaps in healthcare.

Table of Figures

References

[i] Hayes, Adam. "How Multiple Linear Regression Works." *Investopedia*, 21 Sept. 2020, www.investopedia.com/terms/m/mlr.asp.

[ii] Zach. "The Five Assumptions of Multiple Linear Regression." *Statology*, 16 Nov. 2021, www.statology.org/multiple-linear-regression-assumptions/.

[iii] IBM. "What Is Exploratory Data Analysis? | IBM." *Www.ibm.com*, www.ibm.com/topics/exploratory-data-analysis. Accessed 21 Mar. 2023.

[iv] Kenton, Will. "Understanding the Durbin Watson Statistic." *Investopedia*, 2019, www.investopedia.com/terms/d/durbin-watson-statistic.asp.

DECLERATION

I declare that the work I have submitted for this project is my own work and has been completed by me alone. I have acknowledged all material and sources that have been used in its preparation, whether they be books, publications, lecture notes, or any other kind of document. I have not copied or otherwise plagiarised any part of the work submitted for this project from other students and/or persons.

Date: 25/02/2023

Signature: Aleksandra Kolasa