Damilare Kolawole
X21235571

1.0 INTRODUCTION

This study aims to investigate the socio-economic factors influencing the Cancer mortality rate in the US using multiple regression models to help discover socio-economic factor that has an higher influence.

It is important to understand multiple regression and the assumptions associated with it before this research is done.

Simply put, multiple linear regression is a statistical method used to examine the relationship between dependent and several independent variables. To predict the value of the dependent variable, a linear equation is derived using the values of the independent ones.

The different socio-economic factors are the independent variables while the mortality rate is the dependent variable in the US cancer mortality rate study. The following multiple linear regression assumptions should be met.[1]

- Linearity refers to the connection between the dependent variable and the independent variables.
- Independence: The findings should be distinct from one another.
- Homoscedasticity: The residual variance (the gap between predicted and actual values) should be constant across all levels of the independent variables.
- Normality: The residuals should be distributed normally.
- There should be no multicollinearity, which means that the independent variables should not be strongly correlated with one another.

When these assumptions are satisfied, the multiple linear regression model can be constructed by calculating the coefficients of the independent factors in the linear equation. After determining the values of the independent variables, the model can be utilized in forecasting the mortality rate. Various socioeconomic factors associated with cancer can be used as independent variables in the instance of cancer mortality rates in the United States.[2] [3]

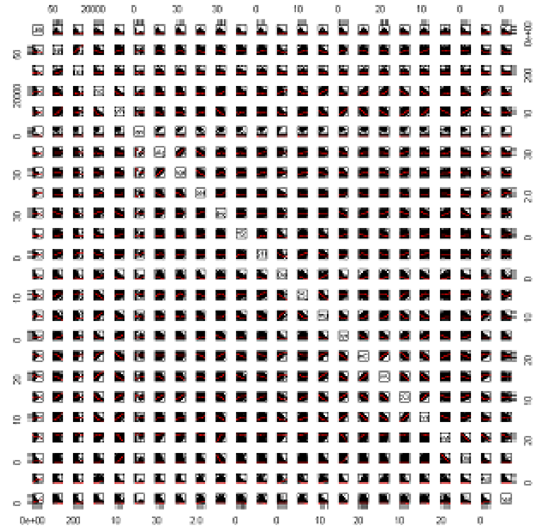
Finally, multiple linear regression is an effective statistical method for examining the relationship between a dependent variable and numerous independent variables. In the case of cancer mortality rates in the United States, different socio-economic factors can be used as independent variables to predict mortality rates. Before conducting multiple linear regression, the assumptions of linearity, and independence, homoscedasticity, normality, and no multicollinearity should be met.

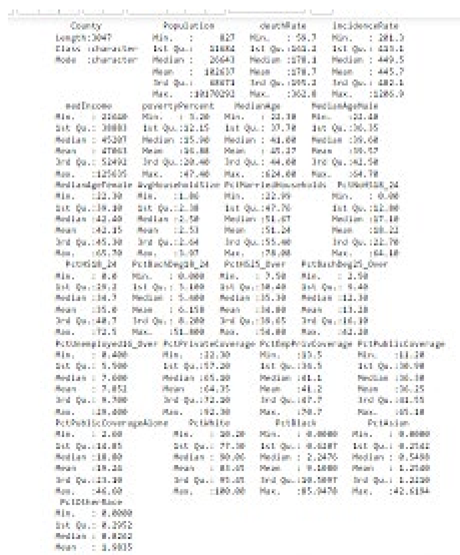
2.0 DESCRIPTION OF DATASET AND ITS VARIABLES

The R programming language will be used to carry out this experiment. The dataset got consists of 3047 observations and 25 columns. Out of these twenty-five columns (25), 23 independent variables and 1 dependent variable. in this case, the use of multiple linear regression becomes more appropriate to analyze the relationship between 2 or more predictor variables and 1 continuous variable. Below are the socio-economic factors.

- Income related: the median income, the percentage of unemployed, and the poverty rate
- Age related: Median Age across the population, and for male and female separately
- Household related: Average Household Size and percentage of Married Households
- Education related: Percentage of the highest educational level attained (No High School / High School / Bachelor Degree) in the age groups 18-24 and over 25.
- Health Insurance related: Percentage of Private Insurance, Private Insurance paid by
- Employer, Public Insurance and Public Insurance Only, and
- Race related: percentage of White/Black/Asian/Other.

County	Population	deathRate	incidenceRate	medIncome	povertyPercent	MedianAge	MedianAgeMale	MedianAgeFemale
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE





```
    County            Population          deathRate      incidenceRate
 Length:3047        Min.   :     827   Min.   : 59.7   Min.   : 201.3
 Class :character   1st Qu.:   11684   1st Qu.:161.2   1st Qu.: 420.3
 Mode  :character   Median :   26643   Median :178.1   Median : 449.5
                    Mean   :  102637   Mean   :178.7   Mean   : 445.7
                    3rd Qu.:   68671   3rd Qu.:195.2   3rd Qu.: 482.1
                    Max.   :10170292   Max.   :362.8   Max.   :1206.9
   medIncome      povertyPercent     MedianAge       MedianAgeMale
 Min.   : 22640   Min.   : 3.20   Min.   : 22.30   Min.   :22.40
 1st Qu.: 38883   1st Qu.:12.15   1st Qu.: 37.70   1st Qu.:36.35
 Median : 45207   Median :15.90   Median : 41.00   Median :39.60
 Mean   : 47063   Mean   :16.88   Mean   : 45.27   Mean   :39.57
 3rd Qu.: 52492   3rd Qu.:20.40   3rd Qu.: 44.00   3rd Qu.:42.50
 Max.   :125635   Max.   :47.40   Max.   :624.00   Max.   :64.70
 MedianAgeFemale AvgHouseholdSize PctMarriedHouseholds  PctNoHS18_24
 Min.   :22.30   Min.   :1.86    Min.   :22.99         Min.   : 0.00
 1st Qu.:39.10   1st Qu.:2.38    1st Qu.:47.76         1st Qu.:12.80
 Median :42.40   Median :2.50    Median :51.67         Median :17.10
 Mean   :42.15   Mean   :2.53    Mean   :51.24         Mean   :18.22
 3rd Qu.:45.30   3rd Qu.:2.64    3rd Qu.:55.40         3rd Qu.:22.70
 Max.   :65.70   Max.   :3.97    Max.   :78.08         Max.   :64.10
  PctHS18_24    PctBachDeg18_24   PctHS25_Over   PctBachDeg25_Over
 Min.   : 0.0   Min.   : 0.000   Min.   : 7.50   Min.   : 2.50
 1st Qu.:29.2   1st Qu.: 3.100   1st Qu.:30.40   1st Qu.: 9.40
 Median :34.7   Median : 5.400   Median :35.30   Median :12.30
 Mean   :35.0   Mean   : 6.158   Mean   :34.80   Mean   :13.28
 3rd Qu.:40.7   3rd Qu.: 8.200   3rd Qu.:39.65   3rd Qu.:16.10
 Max.   :72.5   Max.   :51.800   Max.   :54.80   Max.   :42.20
PctUnemployed16_Over PctPrivateCoverage PctEmpPrivCoverage PctPublicCoverage
 Min.   : 0.400      Min.   :22.30      Min.   :13.5       Min.   :11.20
 1st Qu.: 5.500      1st Qu.:57.20      1st Qu.:34.5       1st Qu.:30.90
 Median : 7.600      Median :65.10      Median :41.1       Median :36.30
 Mean   : 7.852      Mean   :64.35      Mean   :41.2       Mean   :36.25
 3rd Qu.: 9.700      3rd Qu.:72.10      3rd Qu.:47.7       3rd Qu.:41.55
 Max.   :29.400      Max.   :92.30      Max.   :70.7       Max.   :65.10
PctPublicCoverageAlone    PctWhite         PctBlack          PctAsian
 Min.   : 2.60          Min.   : 10.20   Min.   : 0.0000   Min.   : 0.0000
 1st Qu.:14.85          1st Qu.: 77.30   1st Qu.: 0.6207   1st Qu.: 0.2542
 Median :18.80          Median : 90.06   Median : 2.2476   Median : 0.5498
 Mean   :19.24          Mean   : 83.65   Mean   : 9.1080   Mean   : 1.2540
 3rd Qu.:23.10          3rd Qu.: 95.45   3rd Qu.:10.5097   3rd Qu.: 1.2210
 Max.   :46.60          Max.   :100.00   Max.   :85.9470   Max.   :42.6194
 PctOtherRace
 Min.   : 0.0000
 1st Qu.: 0.2952
 Median : 0.8262
 Mean   : 1.9835
 3rd Qu.: 2.1700
```

Figure 3checking if the data is continuous.

A. VISUALIZATION OF THE DATA

In this section, we visualize the data using a boxplot and histogram. Each variable histogram was plotted to see if the data is normalized, and the boxplot function was used to check for outliers in the dataset.

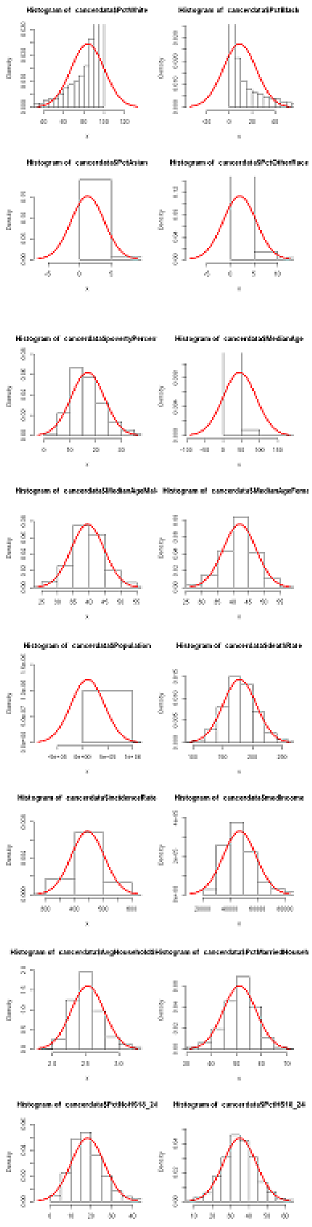


Figure 4Histogram of the various variable

Now to check clearly for the outliers, having done the summary some specific variables were carefully selected. The figure below shows a boxplot of these variables.

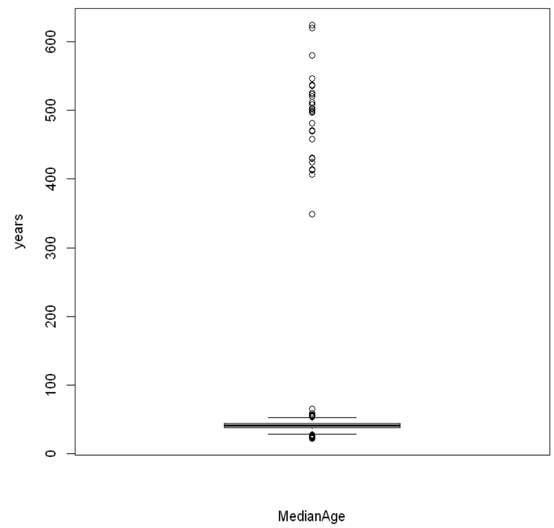


Figure 5 Boxplot for Median Age

Figure 5 shows the boxplot for the median age, we can see so many outliers in the age.

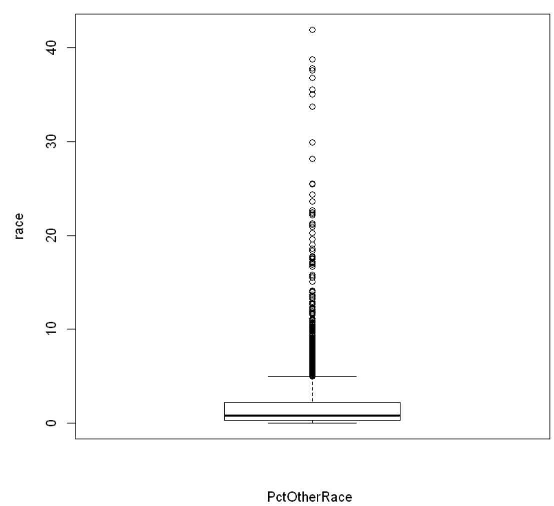


Figure 6 Boxplot for other race

The outliers in other-race variable is also much and must be taken care of and lastly, a boxplot for the percentage of degree bachelors between 18 and 24 is shown in the figure below:

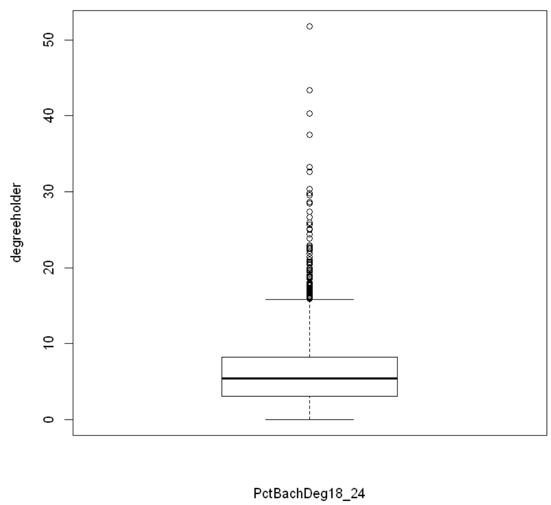


Figure 7 Boxplot for PCTBct18-24

3.0 MODEL

This section discusses the model used and the one which comes out best. After we have processed our data, by checking the missing values of which none was found, also checked if the data is continuous and made sure all assumptions have been met.

Our first model was built by comparing all variables as shown in the figure below:

```
Call:
lm(formula = deathRate ~ Population + incidenceRate + medIncome +
    povertyPercent + MedianAge + MedianAgeMale + MedianAgeFemale +
    AvgHouseholdSize + PctMarriedHouseholds + PctNoHS18_24 +
    PctHS18_24 + PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over +
    PctUnemployed16_Over + PctPrivateCoverage + PctEmpPrivCoverage +
    PctPublicCoverage + PctPublicCoverageAlone + PctWhite + PctBlack +
    PctAsian + PctOtherRace, data = cancerdata)

Residuals:
     Min       1Q   Median       3Q      Max
-109.651  -10.409   -0.145   10.386  138.755

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.759e+02  1.551e+01  11.344  < 2e-16 ***
Population             -1.678e-06  1.237e-06  -1.356 0.175083
incidenceRate           2.049e-01  6.581e-03  31.139  < 2e-16 ***
medIncome               6.616e-05  7.805e-05   0.848 0.396705
povertyPercent          3.671e-01  1.439e-01   2.551 0.010799 *
MedianAge              -2.952e-03  7.627e-03  -0.387 0.698785
MedianAgeMale          -2.195e-01  1.968e-01  -1.115 0.264747
MedianAgeFemale        -2.836e-01  2.158e-01  -1.314 0.188959
AvgHouseholdSize       -1.603e+01  2.710e+00  -5.916 3.68e-09 ***
PctMarriedHouseholds    3.843e-02  9.794e-02   0.392 0.694817
PctNoHS18_24           -8.657e-02  5.450e-02  -1.589 0.112263
PctHS18_24              2.359e-01  4.778e-02   4.938 8.32e-07 ***
PctBachDeg18_24         1.368e-02  1.049e-01   0.130 0.896287
PctHS25_Over            3.227e-01  9.359e-02   3.448 0.000573 ***
PctBachDeg25_Over      -1.246e+00  1.490e-01  -8.366  < 2e-16 ***
PctUnemployed16_Over    4.160e-01  1.569e-01   2.652 0.008055 **
PctPrivateCoverage     -6.753e-01  1.299e-01  -5.197 2.16e-07 ***
PctEmpPrivCoverage      3.714e-01  9.828e-02   3.779 0.000161 ***
PctPublicCoverage      -9.107e-02  2.111e-01  -0.431 0.666269
PctPublicCoverageAlone  2.258e-01  2.661e-01   0.849 0.396174
PctWhite               -1.621e-01  5.691e-02  -2.848 0.004424 **
PctBlack               -7.093e-02  5.387e-02  -1.317 0.188096
PctAsian               -2.711e-02  1.829e-01  -0.148 0.882178
PctOtherRace           -8.786e-01  1.207e-01  -7.279 4.28e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.86 on 3023 degrees of freedom
Multiple R-squared:  0.5416,    Adjusted R-squared:  0.5381
F-statistic: 155.3 on 23 and 3023 DF,  p-value: < 2.2e-16
```

Figure 8: Model 1

The model above shows that there is a relationship between the predictors and the response. Notice that some predictors do not have significant statistical effects. The r squared (R^2) value shows that 53% of the predictors can explain the changes in this multiple regression model.

The next figure below visualizes the model, checking the residuals, normal q-q

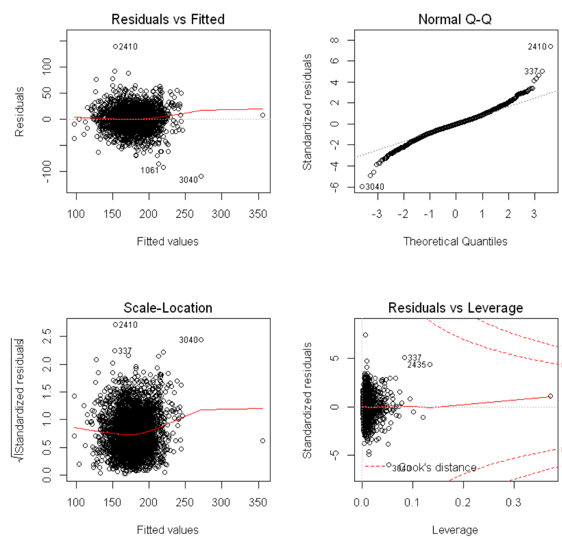


Figure 9 Visualization of Model 1

The figure above shows us visualized results for model 1. We have some outliers that need to be removed before plotting the next model.

```
In [68]: summary(model2)

Call:
lm(formula = deathRate ~ incidenceRate + povertyPercent + AvgHouseholdSize +
    PctHS18_24 + PctHS25_Over + PctBachDeg25_Over + PctUnemployed16_Over +
    PctPrivateCoverage + PctWhite + PctEmpPrivCoverage + PctOtherRace,
    data = cancerdata)

Residuals:
    Min      1Q  Median      3Q     Max
-106.44  -10.63   -0.03   10.47  137.22

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          130.482689  10.855745  12.020  < 2e-16 ***
incidenceRate          0.208603   0.006423  32.478  < 2e-16 ***
povertyPercent         0.539185   0.111106   4.853 1.28e-06 ***
AvgHouseholdSize      -8.906789   1.808379  -4.925 8.87e-07 ***
PctHS18_24             0.224184   0.044376   5.052 4.63e-07 ***
PctHS25_Over           0.244874   0.089816   2.726  0.00644 **
PctBachDeg25_Over     -1.192092   0.132693  -8.984  < 2e-16 ***
PctUnemployed16_Over   0.303681   0.145687   2.084  0.03720 *
PctPrivateCoverage    -0.714072   0.091293  -7.822 7.13e-15 ***
PctWhite              -0.109985   0.028176  -3.904 9.69e-05 ***
PctEmpPrivCoverage     0.546915   0.074722   7.319 3.18e-13 ***
PctOtherRace          -0.832053   0.114478  -7.268 4.61e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.93 on 3035 degrees of freedom
Multiple R-squared:  0.5363,    Adjusted R-squared:  0.5347
F-statistic: 319.2 on 11 and 3035 DF,  p-value: < 2.2e-16
```

Figure 10 Model 2

The model above shows that there is a relationship between the predictors and the response. Notice that some predictors have less significant statistical effects. The r squared (R^2) value shows that 53% of the predictors can explain the changes with an increased F statistics in this multiple regression model.

The next figure below visualizes the model, checking the residuals, normal q-q

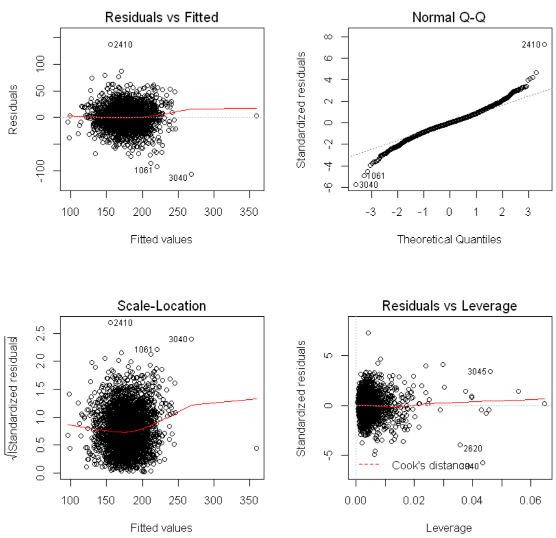


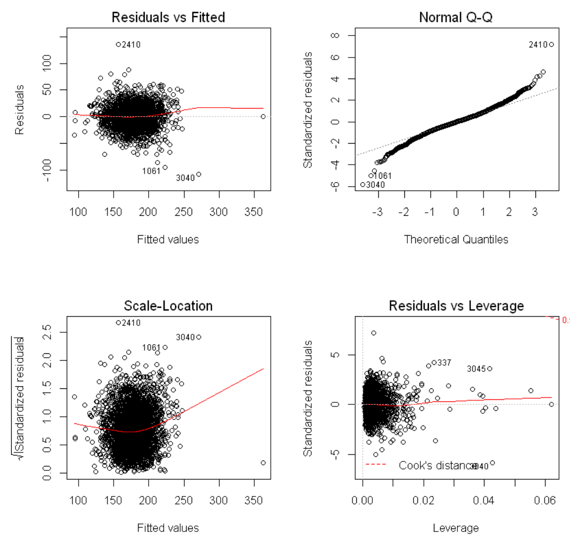
Figure 11 the visualization of model 2

```
Call:
lm(formula = deathRate ~ incidenceRate + povertyPercent + AvgHouseholdSize +
    PctHS18_24 + PctBachDeg25_Over + PctPrivateCoverage + PctWhite +
    PctEmpPrivCoverage + PctOtherRace, data = cancerdata)

Residuals:
     Min       1Q   Median       3Q      Max
-108.387  -10.353    0.125   10.505  135.633

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        144.320794   9.904224  14.572  < 2e-16 ***
incidenceRate        0.211355   0.006328  33.400  < 2e-16 ***
povertyPercent       0.567009   0.109019   5.201 2.11e-07 ***
AvgHouseholdSize    -9.305004   1.769947  -5.257 1.56e-07 ***
PctHS18_24           0.255031   0.043112   5.916 3.68e-09 ***
PctBachDeg25_Over   -1.459241   0.090516 -16.121  < 2e-16 ***
PctPrivateCoverage  -0.735072   0.090252  -8.145 5.50e-16 ***
PctWhite            -0.119089   0.027460  -4.337 1.49e-05 ***
PctEmpPrivCoverage   0.575576   0.074163   7.761 1.14e-14 ***
PctOtherRace        -0.942433   0.109439  -8.611  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.96 on 3037 degrees of freedom
Multiple R-squared:  0.5347,    Adjusted R-squared:  0.5333
F-statistic: 387.8 on 9 and 3037 DF,  p-value: < 2.2e-16
```

Figure 12 Model 3

The model above also shows that there is a relationship between the predictors and the response. In this model, notice that all predictors have significant statistical effects. The r squared (R^2) value shows that 53% of the predictors and also a better F statistics, the residual error was also improved on. This improvement explain the changes in this multiple regression model.

The next figure below visualizes the model, checking the residuals, normal q-q



To see the model more clearly a visualization was done in and it is observed that we have some more outliers that needs to be handled.

The Outliers were all handled, and a new model was plotted. The next figure shows the performance of the new model.

```
Call:
lm(formula = deathRate ~ incidenceRate + povertyPercent + AvgHouseholdSize +
    PctHS18_24 + PctWhite + PctBachDeg25_Over + PctEmpPrivCoverage +
    PctOtherRace, data = cancerdata)

Residuals:
     Min       1Q   Median       3Q      Max
-115.739  -10.526    0.256   10.848  135.181

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        89.902583   7.389240  12.167  < 2e-16 ***
incidenceRate       0.212555   0.006394  33.243  < 2e-16 ***
povertyPercent      1.045378   0.092827  11.262  < 2e-16 ***
AvgHouseholdSize   -3.074254   1.613149  -1.906  0.05678 .
PctHS18_24          0.299996   0.043214   6.942 4.71e-12 ***
PctWhite           -0.124795   0.027744  -4.498 7.12e-06 ***
PctBachDeg25_Over  -1.609662   0.089559 -17.973  < 2e-16 ***
PctEmpPrivCoverage  0.171935   0.055763   3.083  0.00207 **
PctOtherRace       -0.805362   0.109294  -7.369 2.21e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.16 on 3038 degrees of freedom
Multiple R-squared:  0.5245,    Adjusted R-squared:  0.5233
F-statistic: 418.9 on 8 and 3038 DF,  p-value: < 2.2e-16
```

Figure 13 Model 4

The model above shows that there is a relationship between the predictors and the response. But in this model notice that predictors have a very less significant statistical effects. The r squared (R^2) value shows that 52% of the predictors can explain the changes with a better F statistic in this multiple regression model.

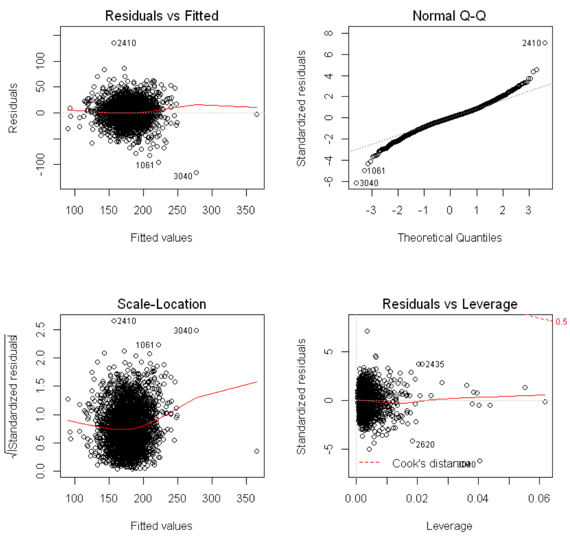


Figure 14 Visualization of Model 4

4.0 FINAL MODEL.

Model 5 was built by removing the less significant predictor.

```
Call:
lm(formula = deathRate ~ incidenceRate + povertyPercent + PctHS18_24 +
    PctWhite + PctBachDeg25_Over + PctEmpPrivCoverage + PctOtherRace,
    data = cancerdata)

Residuals:
    Min      1Q  Median      3Q     Max
-116.25  -10.46    0.25   10.85  135.04

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        81.544163   5.949399  13.706  < 2e-16 ***
incidenceRate       0.214184   0.006339  33.786  < 2e-16 ***
povertyPercent      1.032691   0.092628  11.149  < 2e-16 ***
PctHS18_24          0.292278   0.043043   6.790 1.34e-11 ***
PctWhite           -0.110599   0.026737  -4.137 3.62e-05 ***
PctBachDeg25_Over  -1.594905   0.089262 -17.868  < 2e-16 ***
PctEmpPrivCoverage  0.149609   0.054542   2.743  0.00612 **
PctOtherRace       -0.867645   0.104338  -8.316  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.17 on 3039 degrees of freedom
Multiple R-squared:  0.524,     Adjusted R-squared:  0.5229
F-statistic: 477.9 on 7 and 3039 DF,  p-value: < 2.2e-16
```

Having removed the less significant variable, the model performed better. A better residual standard error was gotten a better F statistics that show how fitted it is and a R-squared value of 52%.

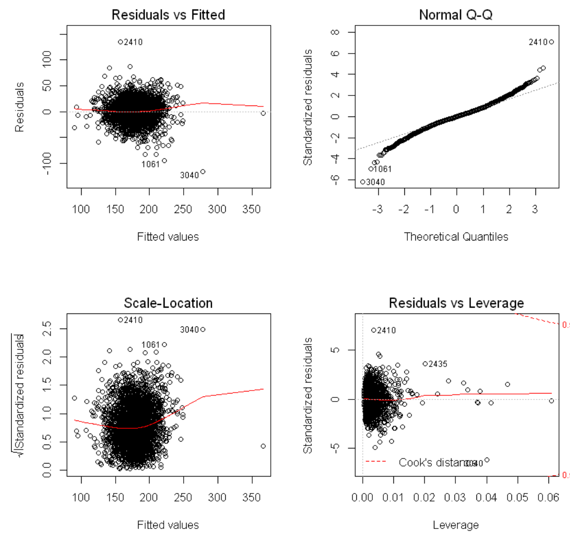


Figure 15 Visualization of the model

In other to avoid overfitting, the outliers seen in the above visualized model would not be considered as all variables are highly significant predictors. For better clarity ANOVA was used to compare all five (5) models.

```
n [84]: anova(model1, model2, model3, model4, model5)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3023	1075265	NA	NA	NA	NA
3035	1087694	-12	-12429.884	2.912116	5.034768e-04
3037	1091528	-2	-3833.722	5.389065	4.610215e-03
3038	1115370	-1	-23841.417	67.027790	3.903041e-16
3039	1116703	-1	-1333.403	3.748730	5.294075e-02

Figure 16 ANOVA Result

5.CONCLUSION

Having ran the test in 5 models, it is observed that the fifth model performed well with a R-squared value of 52% and F-statistics of 477 and a degree of freedom of -1 from the ANOVA comparison.

REFERENCES

[1] C. N. Fru, T. Andrew, F. N. Cho, T. Tassang, and P. N. Fru, "Socio-economic Determinants Influencing Cervical Cancer Screening in Buea: A Cross-Sectional Study," *IJTDH*, pp. 14–22, Aug. 2020, doi: 10.9734/ijtdh/2020/v41i1130331.

[2] T. Akinyemiju, Q. Meng, and N. Vin-Raviv, "Race/ethnicity and socio-economic differences in colorectal cancer surgery outcomes: analysis of the nationwide inpatient sample," *BMC Cancer*, vol. 16, no. 1, p. 715, Dec. 2016, doi: 10.1186/s12885-016-2738-7.

[3] M. Mohebbi, R. Wolfe, D. Jolley, A. B. Forbes, M. Mahmoodi, and R. C. Burton, "The spatial distribution of esophageal and gastric cancer in Caspian region of Iran: An ecological analysis of diet and socio-economic influences," *Int J Health Geogr*, vol. 10, no. 1, p. 13, 2011, doi: 10.1186/1476-072X-10-13.