Employing A Multivariate Linear Regression Model To Predict The Cancer Death Rate Using Several Socioeconomic Characteristics

Pratheek Gogate
National College of Ireland
School of Computing
Dublin, Ireland
X22159789@student.ncirl.ie

***Abstract*—It is wise to find cancer at an early stage utilizing a variety of criteria because it is a very hazardous condition that can be fatal if not treated at the earliest possible stage. So, utilizing the incidence rate and many socioeconomic variables such education, income, insurance, age, household, and race, we were able to estimate the cancer death rate in this study. Because we needed to determine the relationship between the target variable and a number of input factors, multi linear regression was the best option.**

***Index Terms*—component, formatting, style, styling, insert**

I. OBJECTIVES

In today's society, predictions are quite helpful since they allow for the key decisions to be made regardless of the industry because of the predicted outcomes. However, when it comes to the health industry, predictions are even more significant because they have the potential to save many lives. Through the use of multivariate linear regression, we are able to predict the death-rate of cancer patients based on a variety of socioeconomic indicators. This is done in order to determine the link between the output and input variables. Each step of the model construction process uses the Gauss Markov assumptions.

II. DEFINITION AND TERMINOLOGY

A. Descriptive Statistics

A subset of statistics known as descriptive statistics is concerned with gathering, organizing, analyzing, and presenting data in order to highlight and define its key characteristics. To describe the usual value and the variety in the data, it uses measures of central tendency, such as mean, median, and mode, as well as measures of dispersion, such as standard deviation and variance. When comparing diverse datasets and drawing conclusions about bigger populations, descriptive statistics can be used to assist uncover patterns, trends, and relationships within a dataset. In general, it is a helpful tool for comprehending and distilling material in a form that is simple to understand and convey to others.

B. Multiple Linear Regression

A relationship between a single dependent variable and several independent variables can be found statistically using multiple linear regression (MLR). MLR evaluates the impact of several variables on the dependent variable as opposed to ordinary linear regression, which takes only one independent variable into account. The objective of MLR is to produce a precise linear equation that can forecast the dependent variable based on the independent variables. Identification of the pertinent independent variables, model fitting with a regression technique, and multicollinearity testing are the steps in the construction of an MLR model. MLR has several uses in the fields of marketing, finance, and social sciences, including the forecasting of financial trends, the prediction of sales, and the identification of variables affecting customer behavior. The use of MLR is anticipated to increase in the future as data availability rises and algorithms become more sophisticated.

y = 0 + 1x1 + 2x2 + ... + nxn +

where y is the dependent variable, x1, x2,..., xn are the independent variables, 0 is the intercept, 1, 2,..., n are the coefficients indicating the effects of the independent variables on the dependent variable, and is the error term. The MLR equation is used to determine the values of the coefficients 0, 1, 2,..., n that best fit the data and offer insights into the relationships between the variables. The MLR equation is a useful tool for data analysis and decision-making since it can be used to forecast outcomes as well as to pinpoint which independent variables significantly affect the dependent variable.

C. R- Squared

Statistically, the goodness of fit of a regression model is assessed using the R-squared (R2) statistic. The proportion of the dependent variable's variation that the model's independent variables can account for is shown by the R2 value, which has a range of 0 to 1. A high R2 number denotes a strong fit of the model to the data, while a low R2 value denotes a poor match. A regression model's R2 value by itself, however, does not guarantee that the dependent variable will be accurately predicted by it. Consideration should also be given to additional elements like the importance of the individual predictors and the occurrence of multicollinearity. R2, however, is a helpful measure for assessing the overall fit of a regression model.

R2 = 1 - (SSres / SStot)

where: The sum of squares of the residuals (SSres) (the differences between the actual values and the predicted values) The total square sum is denoted by SStot. (the sum of squares of the differences between the actual values and the mean of the dependent variable) R2 is determined as the ratio of the explained variation to the total variance in the dependent variable, to put it another way.

D. Adjusted R - Squared

The adjusted R-squared is a subset of R-squared that accounts for variables in a regression model that are not statistically significant. To put it another way, the adjusted R-squared shows whether or not adding more variables makes a regression model more accurate. R-squared and adjusted R-squared never exceed one another in comparison.

Adjusted R2 = 1 - [(1 - R2) * (n - 1) / (n - k - 1)]

n is the sample size, k is the number of independent variables, and R2 is the coefficient of determination or R-squared value.

E. Residual standard error

The discrepancy between the dependent variable's actual values as observed and its anticipated values as determined by a regression model is known as residual error. Regression analysis relies heavily on this crucial measurement of how well the model fits the data. While a large residual error signifies a bad fit, a little residual error indicates a good fit. Depending on the observed value's relationship to the expected value, residual error might be either positive or negative. To enhance the model's fit, regression analysis seeks to reduce residual error. Residual error is helpful for assessing the effectiveness of regression models, locating outliers and important data points, and identifying other features of the data that may have an impact on the model's performance.

F. Gauss-Markov assumptions

Ordinary least squares (OLS) regression analysis requires the fulfillment of a set of conditions known as the Gauss-Markov assumptions. They comprise:

1.Linearity: The dependent variable's connection to the independent variables must be straight-lined.

2.Independent errors: The residuals or errors must not depend on one another.

3.Homoscedasticity: The variance of the errors must remain the same for all independent variable values.

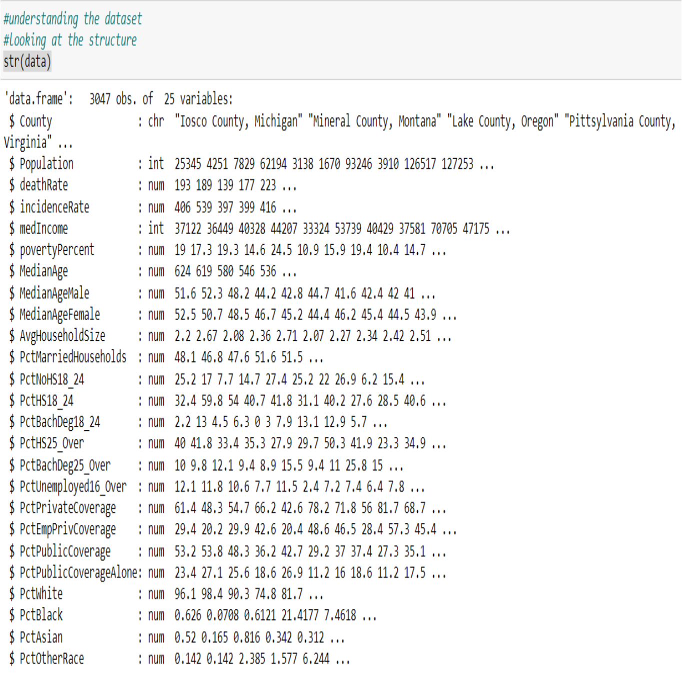
4.Lack of multicollinearity: The independent variables shouldn't have a lot of correlation with one another.

5.Error normality: Errors must have a normal distribution with a mean of zero.

III. DESCRIPTION OF DATA SET

We have a dataset with 3000+ US counties data where our goal is to predict the deathRate based on incidenceRate and different socio economic factors. incidence rate refers to number of newly cancer affected per 100000 people and objective was to find how along with incidence rate, factors like Education, Income, Household, Insurance, Race, Age is going to contribute to the death rate

Target/output variable – deathrate No. of records/rows – 3047 Number of columns – 25 output variable – 1 (deathrate) input variables – 24

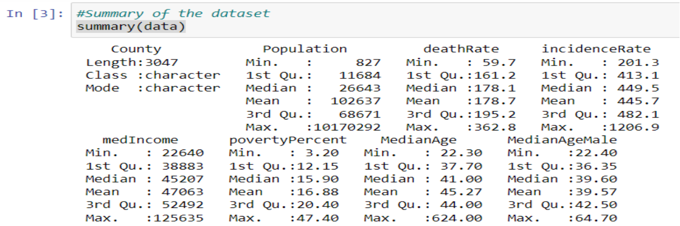


```
#understanding the dataset
#looking at the structure
str(data)
```

```
'data.frame':	3047 obs. of  25 variables:
 $ County                 : chr  "Iosco County, Michigan" "Mineral County, Montana" "Lake County, Oregon" "Pittsylvania County, Virginia" ...
 $ Population             : int  25345 4251 7829 62194 3138 1670 93246 3910 126517 127253 ...
 $ deathRate              : num  193 189 139 177 223 ...
 $ incidenceRate          : num  406 539 397 399 416 ...
 $ medIncome              : int  37122 36449 40328 44207 33324 53739 40429 37581 70705 47175 ...
 $ povertyPercent         : num  19 17.3 19.3 14.6 24.5 10.9 15.9 19.4 10.4 14.7 ...
 $ MedianAge              : num  624 619 580 546 536 ...
 $ MedianAgeMale          : num  51.6 52.3 48.2 44.2 42.8 44.7 41.6 42.4 42 41 ...
 $ MedianAgeFemale        : num  52.5 50.7 48.5 46.7 45.2 44.4 46.2 45.4 44.5 43.9 ...
 $ AvgHouseholdSize       : num  2.2 2.67 2.08 2.36 2.71 2.07 2.27 2.34 2.42 2.51 ...
 $ PctMarriedHouseholds   : num  48.1 46.8 47.6 51.6 51.5 ...
 $ PctNoHS18_24           : num  25.2 17 7.7 14.7 27.4 25.2 22 26.9 6.2 15.4 ...
 $ PctHS18_24             : num  32.4 59.8 54 40.7 41.8 31.1 40.2 27.6 28.5 40.6 ...
 $ PctBachDeg18_24        : num  2.2 13 4.5 6.3 0 3 7.9 13.1 12.9 5.7 ...
 $ PctHS25_Over           : num  40 41.8 33.4 35.3 27.9 29.7 50.3 41.9 23.3 34.9 ...
 $ PctBachDeg25_Over      : num  10 9.8 12.1 9.4 8.9 15.5 9.4 11 25.8 15 ...
 $ PctUnemployed16_Over   : num  12.1 11.8 10.6 7.7 11.5 2.4 7.2 7.4 6.4 7.8 ...
 $ PctPrivateCoverage     : num  61.4 48.3 54.7 66.2 42.6 78.2 71.8 56 81.7 68.7 ...
 $ PctEmpPrivCoverage     : num  29.4 20.2 29.9 42.6 20.4 48.6 46.5 28.4 57.3 45.4 ...
 $ PctPublicCoverage      : num  53.2 53.8 48.3 36.2 42.7 29.2 37 37.4 27.3 35.1 ...
 $ PctPublicCoverageAlone: num  23.4 27.1 25.6 18.6 26.9 11.2 16 18.6 11.2 17.5 ...
 $ PctWhite               : num  96.1 98.4 90.3 74.8 81.7 ...
 $ PctBlack               : num  0.626 0.0708 0.6121 21.4177 7.4618 ...
 $ PctAsian               : num  0.52 0.165 0.816 0.342 0.312 ...
 $ PctOtherRace           : num  0.142 0.142 2.385 1.577 6.244 ...
```

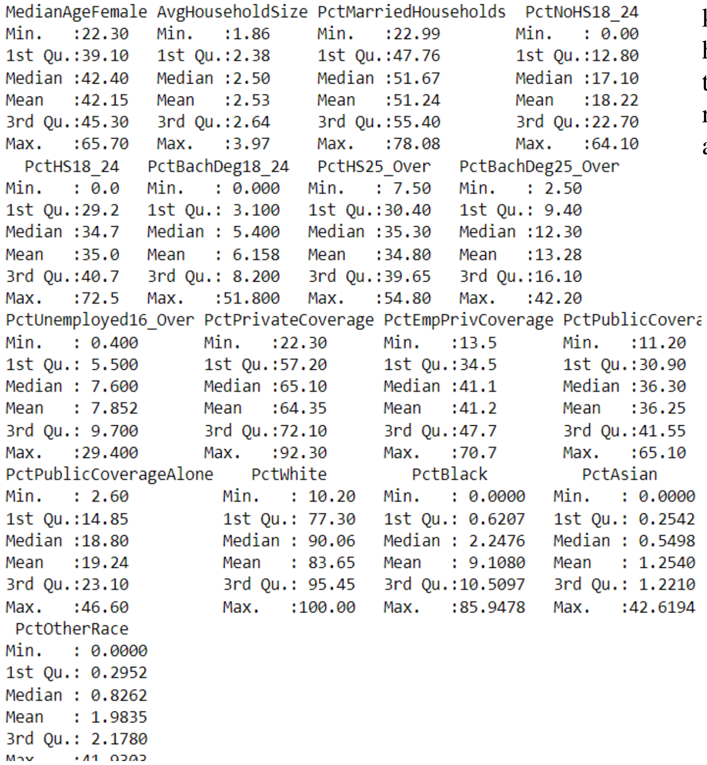
Fig. 1. Image shows the structure of the data

The command "Summary" is very helpful because it provides a wealth of information about the data, including the maximum and minimum values we can use to determine the data's range and the mean and median, both of which are quite useful. If the difference between the mean and median is greater, it suggests that there are outliers; in our case, the Population column has this issue, but we can ignore it because it has no bearing on the outcome that will be revealed in the future steps. Descriptive statistics aid in interpreting the data more fully, and this Summary essentially covers that portion.



```
In [3]: #Summary of the dataset
        summary(data)
```

```
    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.: 413.1
 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 PctPublicCovera
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.   :85.9478   Max.   :42.6194
 PctOtherRace
Min.   : 0.0000
1st Qu.: 0.2952
Median : 0.8262
Mean   : 1.9835
3rd Qu.: 2.1780
Max.   :41.9303
```

Fig. 2. Image shows the summary of the data

IV. DATA VISUALIZATION

Visualization is important in all forms of data. It mainly involves creating visual graphs and charts that help us understand and draw conclusions from the data presented. Hence, in our situation, we utilized numerous plots to check for any outliers and we also plotted graphs for each socioeconomic component against the Deathrate, which helped us to identify the direction of the Deathrate based on the direction of socio variables.

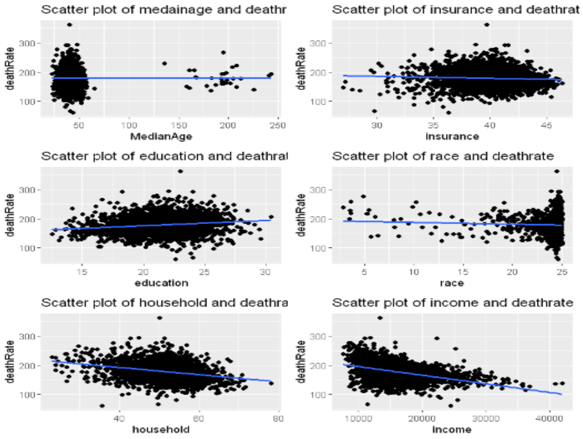


Fig. 3. scatter plots for each socio economic factor against deathrate

Since the points are dispersed and the direction of the lines is negative, which is actually incorrect given the domain knowledge, it is easy to see that almost all of the variables have very weak correlations with death rates. When it comes to race and insurance, there doesn't appear to be much of a relationship because we can't see any rise or fall in death rates as a result of that, which may be true.

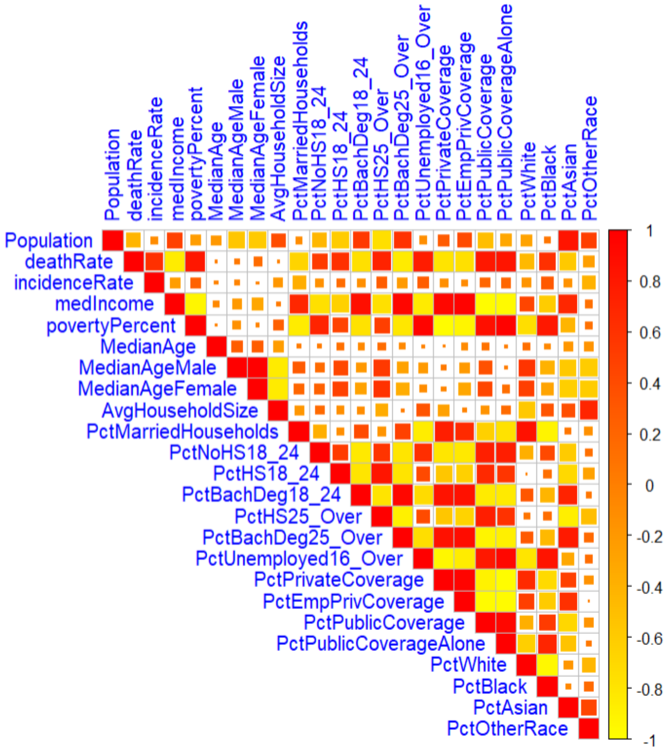
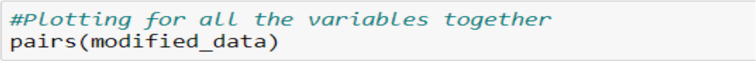


Fig. 4. correlation diagram



```
#Plotting for all the variables together
pairs(modified_data)
```

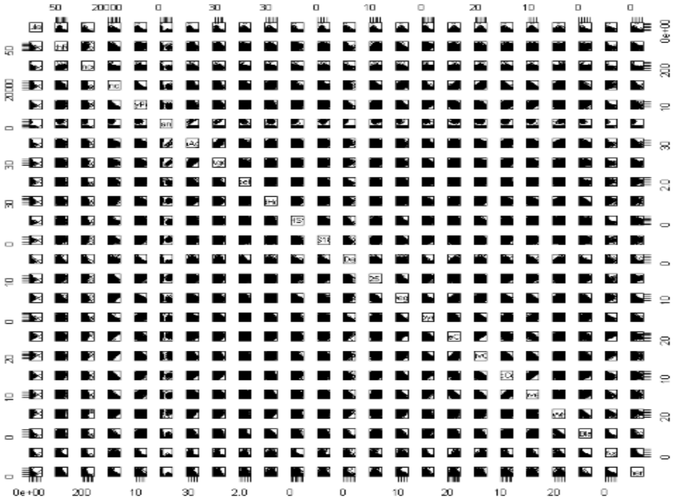


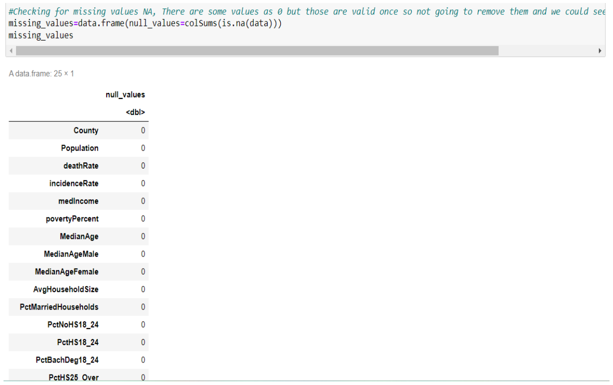
Fig. 5. scatter diagram against all the inputs

Correlation is something that is positive if it occurs between input and output variables and negative when it occurs between input variables. There shouldn't be a lot of collinearity between two input variables because this indicates that there is a relationship between them and that changing one will change the other as well, which will ultimately impact the output variable.

We deduced from the aforementioned graph and the code[1][3] we wrote for correlation that there is collinearity between distinct columns, but since they all belong to the same socioeconomic component, they can be ignored, and income is also tangentially connected with insurances of all kinds.

V. Data Validation

We will verify the data under data validation to see if there are any records with null values.



```
#Checking for missing values NA, There are some values as 0 but those are valid once so not going to remove them and we could see
missing_values=data.frame(null_values=colSums(is.na(data)))
missing_values
```

A data.frame: 25 × 1

	null_values
	<dbl>
County	0
Population	0
deathRate	0
incidenceRate	0
medIncome	0
povertyPercent	0
MedianAge	0
MedianAgeMale	0
MedianAgeFemale	0
AvgHouseholdSize	0
PctMarriedHouseholds	0
PctNoHS18_24	0
PctHS18_24	0
PctBachDeg18_24	0
PctHS25_Over	0

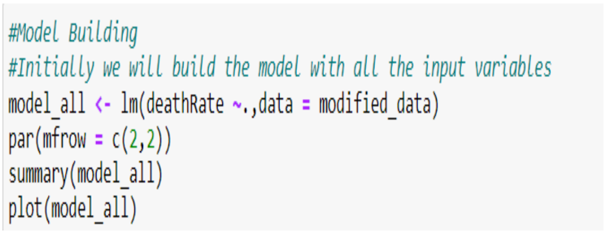
Fig. 6. Finding Null Values

VI. Model Building

In this section of the report, we will provide a quick overview of the steps followed to create a multiple linear regression model.

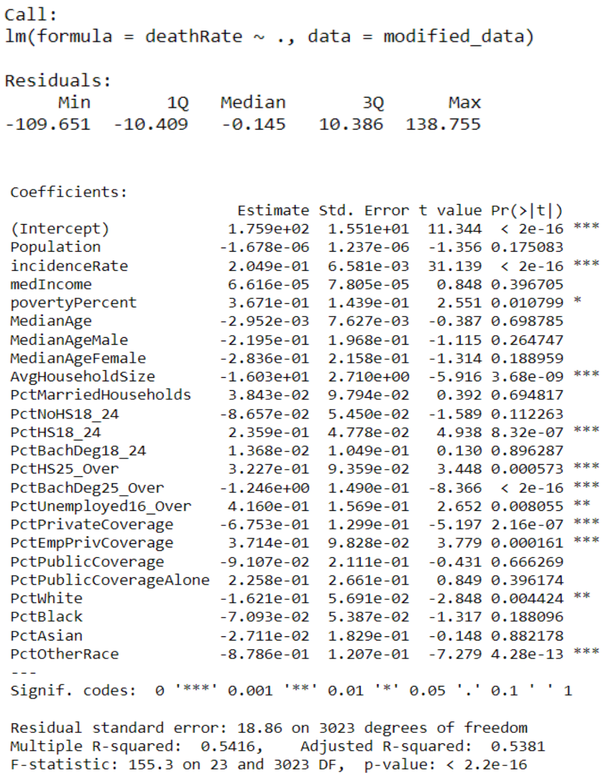
A. Model 1

Consequently, after performing exploratory data analysis, creating numerous graphs, and observing the data, we decided to omit the county column because it has no bearing on the output. Instead, we took into account all the other input variables to build a model in order to determine the relative importance of each variable and to make future decisions regarding which model provides the best case.So for now we are building the model deathrate with all other input variables.



```
#Model Building
#Initially we will build the model with all the input variables
model_all <- lm(deathRate ~.,data = modified_data)
par(mfrow = c(2,2))
summary(model_all)
plot(model_all)
```

Fig. 7. Code for Model 1



```
Call:
lm(formula = deathRate ~ ., data = modified_data)

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
```

Fig. 8. output for Model 1

The output of the model we just created is depicted in the above figure, and it is obvious that there are many non-essential variables present. The last column contains stars that indicate the significance of the variables based on the number of stars; three stars indicate maximum significance, while no stars indicate no significance. As a result, we can remove all those columns from the next model. Looking at the output and results, we can see that the r square value is 0.54—which is a little low—and the residual error is around 10—which isn't too bad but that we could potentially improve. The residual vs.

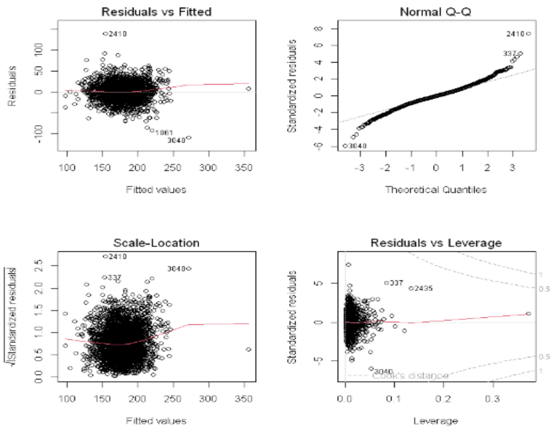
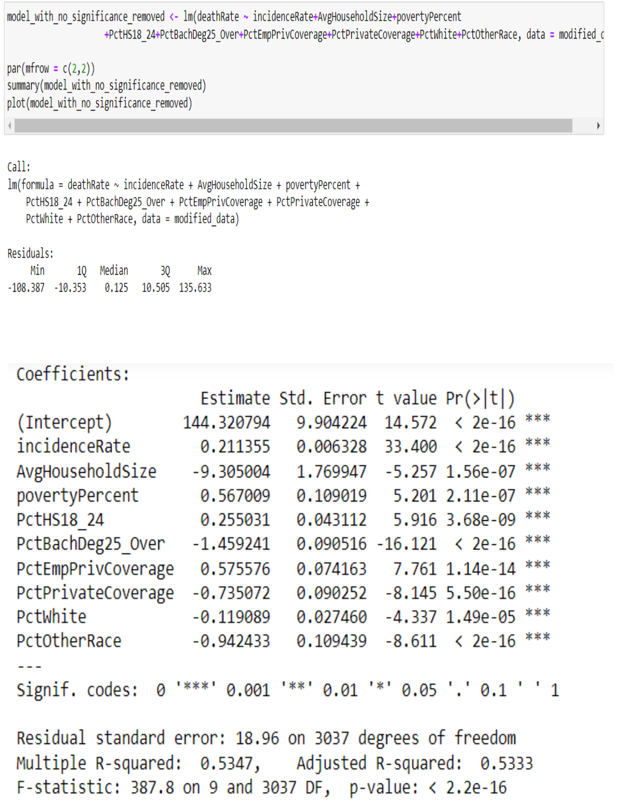


Fig. 9. output graph for Model 1

fitted graph also reveals a small amount of curve that shouldn't be there, and the fourth graph displays some outliers, both of which indicate that we can do better.

B. Model 2

Will try to rectify the issues we had in the first model by firstly eliminating all the unncessary columns.



```
model_with_no_significance_removed <- lm(deathRate ~ incidenceRate+AvgHouseholdSize+povertyPercent
                    +PctHS18_24+PctBachDeg25_Over+PctEmpPrivCoverage+PctPrivateCoverage+PctWhite+PctOtherRace, data = modified_c

par(mfrow = c(2,2))
summary(model_with_no_significance_removed)
plot(model_with_no_significance_removed)
```

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

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 ***
AvgHouseholdSize    -9.305004   1.769947  -5.257 1.56e-07 ***
povertyPercent       0.567009   0.109019   5.201 2.11e-07 ***
PctHS18_24           0.255031   0.043112   5.916 3.68e-09 ***
PctBachDeg25_Over   -1.459241   0.090516 -16.121  < 2e-16 ***
PctEmpPrivCoverage   0.575576   0.074163   7.761 1.14e-14 ***
PctPrivateCoverage  -0.735072   0.090252  -8.145 5.50e-16 ***
PctWhite            -0.119089   0.027460  -4.337 1.49e-05 ***
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
```

Fig. 10. code and output for Model 2

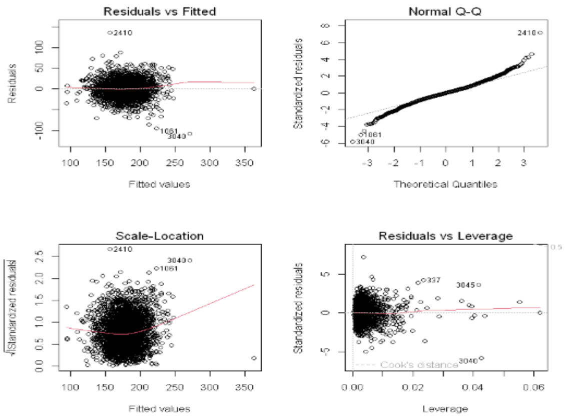
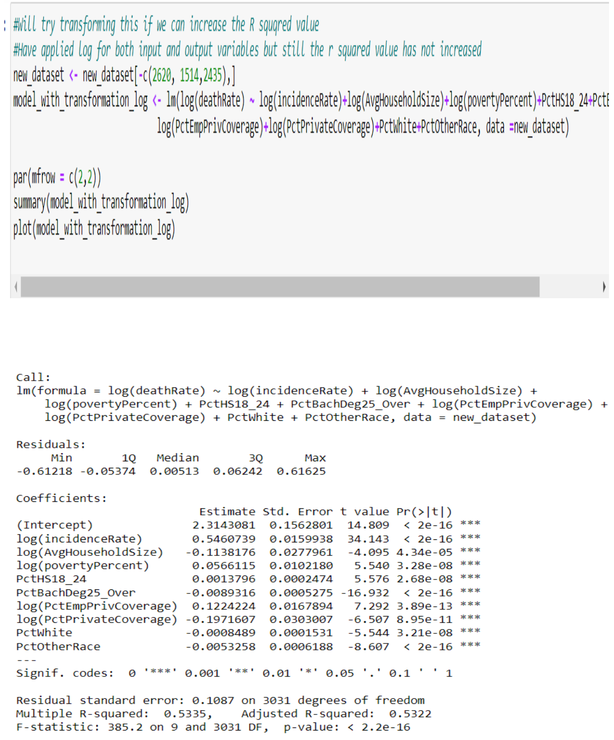


Fig. 11. output graph for Model 2

We therefore created the model using only the necessary essential variables, and when we look at the output, we see no improvement at all, as the residual error is further increased and the r squared value is also unchanged. If we also take a look at the graph, it is linearity-wise okay, but when it comes to the residual vs fitted line, it is in curve shape even though the points appear to be equally distributed on both sides and, in comparison, there are fewer values away from the line.

C. Model 3

Therefore, after looking at the two models we built above, we could see that there was little improvement even after removing the non-essential variables. We also tried building the models with individual socioeconomic factors, but we still did not get the expected r square value, so we had to transform the variables for higher accuracy. We did this by applying log and sqrt for different variables in different combinations, and the best value was obtained using the following combination:



```
#Will try transforming this if we can increase the R squqred value
#Have applied log for both input and output variables but still the r squared value has not increased
new_dataset <- new_dataset[-c(2620, 1514,2435),]
model_with_transformation_log <- lm(log(deathRate) ~ log(incidenceRate)+log(AvgHouseholdSize)+log(povertyPercent)+PctHS18_24+PctE
                                    log(PctEmpPrivCoverage)+log(PctPrivateCoverage)+PctWhite+PctOtherRace, data =new_dataset)


par(mfrow = c(2,2))
summary(model_with_transformation_log)
plot(model_with_transformation_log)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.61218 -0.05374  0.00513  0.06242  0.61625

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.3143081  0.1562801  14.809  < 2e-16 ***
log(incidenceRate)       0.5460739  0.0159938  34.143  < 2e-16 ***
log(AvgHouseholdSize)   -0.1138176  0.0277961  -4.095 4.34e-05 ***
log(povertyPercent)      0.0566115  0.0102180   5.540 3.28e-08 ***
PctHS18_24               0.0013796  0.0002474   5.576 2.68e-08 ***
PctBachDeg25_Over       -0.0089316  0.0005275 -16.932  < 2e-16 ***
log(PctEmpPrivCoverage)  0.1224224  0.0167894   7.292 3.89e-13 ***
log(PctPrivateCoverage) -0.1971607  0.0303007  -6.507 8.95e-11 ***
PctWhite                -0.0008489  0.0001531  -5.544 3.21e-08 ***
PctOtherRace            -0.0053258  0.0006188  -8.607  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1087 on 3031 degrees of freedom
Multiple R-squared:  0.5335,    Adjusted R-squared:  0.5322
F-statistic: 385.2 on 9 and 3031 DF,  p-value: < 2.2e-16
```

Fig. 12. code and output for Model 3

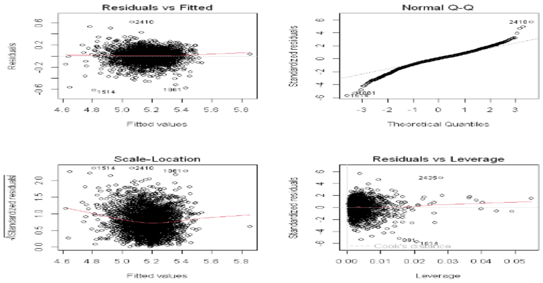


Fig. 13. output graph for Model 3

Compare to all the models we built this was the best as it has a residual error of 0.10 and looking at the graphs first graph looks better tha all the previous models and even though there is a curve in graph 3 the points looks evenly distributed on both the sides of line which indicates there is Homoscedasticity and in the graph two we could see almost all points fall on line which indicates that error is normally distributed.

VII. GAUSS MARKOV ASSUMPTIONS: DIAGNOSIS METHODS

The Ordinary Least Squares (OLS) estimator is the Best Linear Unbiased Estimator (BLUE) of the regression coefficients if the Gauss-Markov assumptions are satisfied. These assumptions include:

A. Linearity

The characteristic of a system or function known as linearity is the fact that the output is inversely proportional to the input. To put it another way, if the input is doubled, the output will also be doubled. The importance of this property can be seen in a number of branches of mathematics, physics, engineering, and economics. In linear systems, the concept of superposition holds, which means that the combined effect of two inputs is equal to the sum of their individual effects. Systems can be more readily understood and controlled thanks to the mathematical analysis that is sped up by linearity. On the other hand, nonlinear systems might be significantly more difficult to evaluate because they do not show the same proportionality between input and output.

B. Homoscedasticity

When a set of data exhibits homoscedasticity, the variance of the residuals, or differences between the actual values observed and those predicted by a statistical model, is constant at all levels of the independent variable (s). In other words, it refers to a situation where the distribution of the data points is constant over the predictor variable's range of values. Several statistical tests, including regression analysis, rely heavily on the premise of homoscedasticity, and failure to meet this requirement can provide findings that are skewed and untrustworthy.

C. Normally distributed errors

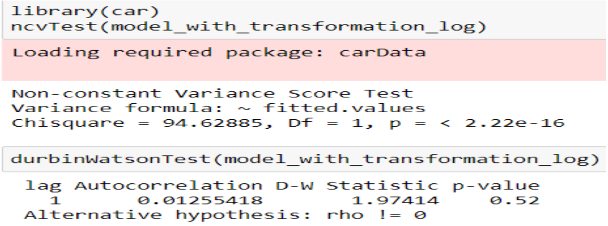
In statistical modeling, it's usual to assume that errors are normally distributed, commonly referred to as Gaussian errors. The assumption of regularly distributed errors indicates that the mistakes in a model follow a normal distribution with a mean of zero and a constant variance. Because it enables the application of robust statistical techniques like maximum likelihood estimation and hypothesis testing, this assumption is crucial. Moreover, naturally occurring events frequently show regularly distributed mistakes, making them an acceptable assumption in many situations. The results of the statistical analysis could be prejudiced or deceptive if the assumption of normally distributed errors is broken.

D. No influential data

No impactful data points refer to observations that do not significantly affect the findings of a statistical study. The elimination of these data points does not significantly alter the conclusions that were reached from the data and has minimal bearing on the overall result. Identifying and correcting significant data points is vital to ensure the correctness and validity of statistical analysis.

E. No autocorrelation between errors

We employ the Durbin-Watson test to determine whether there is no autocorrelation between the errors. The statistic's value must fall between 1 and 3. This is the best approach to comprehend the test. Below is a screenshot of our third model being tested and we could see that value is well inside that range. The NCV test for the model can also be run, and the outcomes can be examined. The NCV (nerve conduction velocity) calculation code[2] is as shown in the above fig, the P value should be as less as possible and in our case which is true.



```
library(car)
ncvTest(model_with_transformation_log)

Loading required package: carData

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 94.62885, Df = 1, p = < 2.22e-16

durbinWatsonTest(model_with_transformation_log)

 lag Autocorrelation D-W Statistic p-value
   1      0.01255418       1.97414    0.52
 Alternative hypothesis: rho != 0
```

Fig. 14. Test Results for Model3

VIII. SUMMARY OF THE MODEL

First, we imported the dataset and tried learning more about the data to understand what the data is about and what it contains. We first checked for missing values, then tried searching for patterns in the data by plotting various graphs, then tried determining correlation between variables by using scatter plots and other methods, and finally built a model using all the variables against Deathrate. In the subsequent model, we eliminated all the unnecessary variables, and we then built a model, but it was still not very accurate so we delited some outliers and also applied log as a tr. Prior to that, we constructed models based on various socioeconomic variables and numerous other combinations, but out of all of those, the one with log had the best result possible because the error was 0.10 even though the R square value was around 54. As a result, after conducting tests and looking at various values, we decided that this was the best model.

IX. REFERENCES

[1] NCI Moodle:https://mymoodle.ncirl.ie/course/view.php?id=1593

[2] ExcelR Data Science course Bangalore

[3] W3 Schools:https://www.w3schools.com/r/