**Abstract**

Using cross sectional data for 2809 counties in the U.S., this paper empirically examines the mortality rate of cancer through different socioeconomic independents. The goal of this project will be to predict this target variable based on six different features that cover demographics and healthcare related information. The result indicates that socioeconomic features, like financial access to healthcare and the income level of its inhabitants highly impact a county’s number of fatal cases resulting from this disease.

**Introduction**:

It is no surprised that the disparity in cancer incidence and mortality rates between high income and low-income areas is wide and increasing. It is important to know that in the US much like many other countries, low-income counties tend to suffer from less education and food insecurity which leads people to practice unsafe health habits including relying on unhealthy foods and low exercise levels. More so, the lack of availability/trust of the American healthcare system in these areas have made it hard to increase the quality of care that patients receive.

Often, it is not a matter of amount of dollars in the pocket, but the poverty level of a county that indicate the level of good health habits and the importance frequent checks, not only that, but it also points to how much of the country’s quality healthcare is allocated to it. The median income also increases the number of people getting health insurance which encourages to get regular checks and have access to higher quality hospitals and professionals.

A broad disparity in lung cancer incidence and mortality, the most common in cancer cases, is highlighted between high- and low-income counties, which confirms that income level and bad health habits such as smoking and poor diet are correlated.

**Data**

The impact of socioeconomic impact on cancer is conducted less frequently because of the unavailability of such data since a person’s economic status is not recorded at the time of death, fortunately, the US federal government has published open data fitting the present need. Cancer.gov website can be used for cancer incidence and mortality data, and the Census American Community Survey for additional variables that were chosen to be included in the model, such as Poverty Rate, Median Income, Insurance Coverage, Population Estimate. The first two datasets (incidence and mortality) were loaded from the cancer.gov page and the others (poverty, income, and health coverage) were made from concatenated tables from the U.S. Census Bureau. They were later merged together to form a dataset containing all the variables mentioned previously with detailed columns by gender and race.

Some low population counties reported less than 16 deaths and therefore did not have information on the mortality rate (it was left blank), so it was decided to remove such records because low population areas are well represented in the dataset and it is actually expected that low population counties have more deaths per capita than larger ones so it was considered that they rather failed to record cancer deaths.

**Study design:**

The study methodology depends on employing a quantitative analysis of secondary data.

1**. Descriptive statistics**: They provide simple summaries about the sample and the measures. They are the appropriate tests in order to address the study questions.

2**. Correlation analysis**: This particular type of analysis is useful when a studier wants to establish if there are possible connections between variables. To test the correlation between 2 variables a bivariate test called Pearson Product Moment Correlation Coefficient is used

3**. Multiple Linear Regression Model assumptions**: According to Saunders et al., (2009), there are 3 main regression models’ assumptions:

a**. Homoscedasticity**: it occurs when the variance of error term is constant across the number of observations (Saunders et al., 2009) that can be seen through residual/predicted regression scatter plot. In order to overcome that problem if it occurs, heteroscedasticity software developed by Andrew Hayes that uses many heteroscedasticity tests already developed by white/Huber or Davindson can be used.

b**. Absence of Multicollinearity**: it occurs when variables are related. To detect multicollinearity, the study has used a Tolerance test. If multicollinearity exists between 2 variables, then one of them should be omitted.

**c. Normality**: it occurs when error values are normally distributed between observed and predicted ones.

**4. Regression analysis**: it is used to assess the relationship between the dependent and independent variables. It is investigated using R by splitting the sample into 80% training data and 20% for validation. Various statistics and tests will be considered namely the p-value and the R-square to asses if the model is overall significant. A p-value < 0.05 indicates that a variable or the entire model is statistically significant and the R-square gives an idea about the percentage of variance in the target variable explained by the model.

**5. Goodness of fit of model**: A statistical model explains how well it fits a set of observations.

6**. Internal cross validation**: A statistical method to explain how well the chosen best fitted model explains the main internal data.

**Hypothesis**:

H0: There is a linear relationship between socioeconomic status and the mortality rate in cancer cases.

H1: There is a no linear relationship between socioeconomic status and the mortality rate in cancer cases.

**Descriptive analysis:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | *Poverty\_PC* | *Median\_Income* | *Health Insurance* | *No Health Insurance* | *Incidence\_Rate* | |  |  |  |  |  |  | | Mean | 16251.7367 | 46812.8904 | 84378.6724 | 12764.0802 | 70.1744037 | | Standard Error | 115.45683 | 234.344767 | 120.126843 | 90.6805125 | 0.32091733 | | Median | 15640.8734 | 45048 | 85312.8542 | 12380.3517 | 69.7 | | Standard Deviation | 6119.21201 | 12420.2727 | 6366.72269 | 4806.06716 | 17.0086185 | | Sample Variance | 37444755.6 | 154263173 | 40535157.8 | 23098281.6 | 289.293102 | | Kurtosis | 1.15774248 | 3.59624269 | 3.01052528 | 1.71288723 | 3.15844978 | | Skewness | 0.81121846 | 1.3794631 | -1.2418852 | 0.84056854 | 0.68590877 | | Range | 43891.3248 | 104125 | 52813.2177 | 39401.9634 | 190.2 | | Minimum | 2598.61791 | 19328 | 43907.377 | 2160.56079 | 13.5 | | Maximum | 46489.9427 | 123453 | 96720.5947 | 41562.5242 | 203.7 | |  |  |  |  |  | *Population\_estimate* | *Falling* | *Rising* | *Poverty* | *Mortality\_Rate* |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | 113930.308 | 0.07013172 | 0.01388394 | 16921.79067 | 53.18853685 |
|  |  |  |  |  |  | 6534.70423 | 0.00481914 | 0.00220812 | 1083.053918 | 0.264944844 |
|  |  |  |  |  |  | 31183 | 0 | 0 | 5020 | 52.6 |
|  |  |  |  |  |  | 17403 | 0 | 0 | 1924 | 48.3 |
|  |  |  |  |  |  | 346339.324 | 0.25541434 | 0.11703015 | 57401.85766 | 14.04207671 |
|  |  |  |  |  |  | 1.1995E+11 | 0.06523649 | 0.01369606 | 3294973263 | 197.1799184 |
|  |  |  |  |  |  | 297.211224 | 9.35307973 | 67.1613428 | 380.4419559 | 1.289549329 |
|  |  |  |  |  |  | 13.3574661 | 3.36844487 | 8.31345443 | 15.54614204 | 0.550068197 |
|  |  |  |  |  |  | 10167990 | 1 | 1 | 1800086 | 116.4 |
|  |  |  |  |  |  | 2302 | 0 | 0 | 179 | 9.2 |
|  |  |  |  |  |  | 10170292 | 1 | 1 | 1800265 | 125.6 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Population\_estimate* | *Falling* | *Rising* | *Poverty* | *Mortality\_Rate* |
|  |  |  |  |  |
| 113930.308 | 0.07013172 | 0.01388394 | 16921.79067 | 53.18853685 |
| 6534.70423 | 0.00481914 | 0.00220812 | 1083.053918 | 0.264944844 |
| 31183 | 0 | 0 | 5020 | 52.6 |
| 346339.324 | 0.25541434 | 0.11703015 | 57401.85766 | 14.04207671 |
| 1.1995E+11 | 0.06523649 | 0.01369606 | 3294973263 | 197.1799184 |
| 297.211224 | 9.35307973 | 67.1613428 | 380.4419559 | 1.289549329 |
| 13.3574661 | 3.36844487 | 8.31345443 | 15.54614204 | 0.550068197 |
| 10167990 | 1 | 1 | 1800086 | 116.4 |
| 2302 | 0 | 0 | 179 | 9.2 |
| 10170292 | 1 | 1 | 1800265 | 125.6 |
|  |  |  |  |  |

Some of the variables described above have a kurtosis far from 1 indicating that the distribution is not normal which means that it is better to transform them using the log or square root function in order to get a better model.

**Correlations:**

A strong pearson correlation can be quickly spotted between “population estimate” and the “poverty rate” so one of them will be removed, this is also why the “poverty per capita” variable was added. A strong negative correlation is observed between “health insurance” and “no health insurance”, the reason why both are included is to test which one has a higher significance when fitting the model.

Another strong correlation is between “incidence” and “mortality rate” which emphasizes its significance. The “Median income” variable correlates with the “Poverty\_PC” at a pearson correlation of 0.78 and both moderately correlated with the “mortality\_rate”. Further analysis will show which data points to keep in order to avoid multicollinearity.

**Preliminary Models:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***Poverty\_PC*** | ***Median\_Income*** | ***Health Insurance*** | ***No Health Insurance*** | ***Incidence\_Rate*** | ***Population\_estimate*** | ***Falling*** | ***Rising*** | ***All\_Poverty*** | ***Mortality\_Rate*** |
| Poverty\_PC | 1 |  |  |  |  |  |  |  |  |  |
| Median\_Income | -0.783066291 | 1 |  |  |  |  |  |  |  |  |
| Health Insurance | -0.34813708 | 0.312608776 | 1 |  |  |  |  |  |  |  |
| No Health Insurance | 0.569161133 | -0.452897992 | -0.733631339 | 1 |  |  |  |  |  |  |
| Incidence\_Rate | 0.316603833 | -0.370311422 | -0.009165488 | 0.036224879 | 1 |  |  |  |  |  |
| Population\_estimate | -0.075232993 | 0.253717194 | -0.001556621 | -0.013322505 | -0.15510346 | 1 |  |  |  |  |
| Falling | -0.057651006 | 0.150961645 | 0.024888753 | -0.025174493 | -0.079259452 | 0.297759261 | 1 |  |  |  |
| Rising | 0.019051288 | -0.01770224 | 0.042151225 | -0.020617684 | -0.009947742 | -0.023751822 | -0.032587 | 1 |  |  |
| All\_Poverty | 0.021132267 | 0.127667228 | -0.041113567 | 0.046900291 | -0.126108073 | 0.968312646 | 0.26508 | -0.0212 | 1 |  |
| Mortality\_Rate | **0.39787861** | **-0.447109979** | **-0.09531413** | **0.161757568** | 0.842383528 | -0.1844818 | -0.068925 | -0.0154 | -0.145648659 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ***Incidence\_Rate*** | ***Population\_estimate*** | ***Falling*** | ***Rising*** | ***All\_Poverty*** | ***Mortality\_Rate*** |
| Incidence\_Rate | 1 |  |  |  |  |  |
| Population\_estimate | -0.15510346 | 1 |  |  |  |  |
| Falling | -0.079259452 | 0.297759261 | 1 |  |  |  |
| Rising | -0.009947742 | -0.023751822 | -0.032587 | 1 |  |  |
| All\_Poverty | -0.126108073 | 0.968312646 | 0.26508 | -0.0212 | 1 |  |
| Mortality\_Rate | **0.842383528** | **-0.1844818** | **-0.068925** | **-0.0154** | **-0.145648659** | 1 |

Every variable in this dataset is significant when paired with the dependent variable “Mortality\_Rate” except for “Rising” indicating that the rising recent trend in cancer cases do not influence the mortality rate.

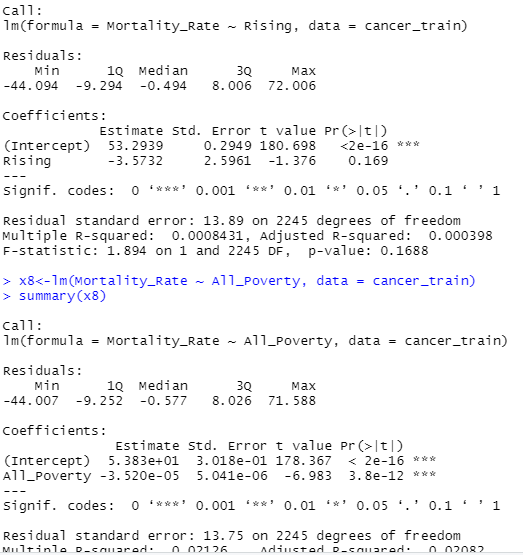
However, the Falling trend does, therefore it seems that when a county decides to take cancer prevention measures, a decrease in the mortality rate might occur since diagnosis is being done earlier.

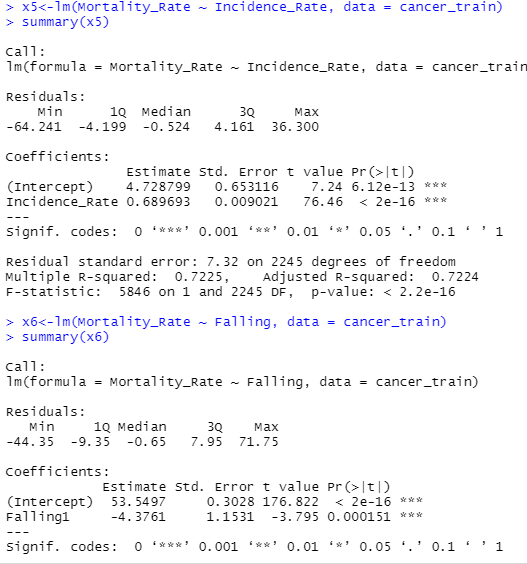
The Incidence Rate increases the Mortality Rate by 0.68 to 1.

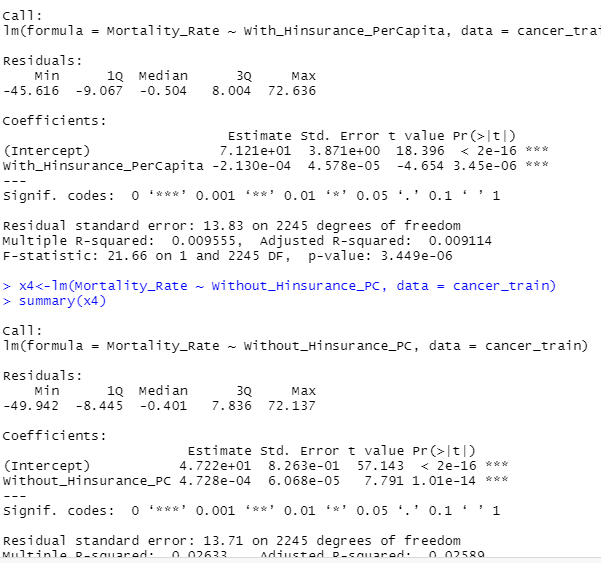
A higher median income decreases cancer fatality probably because of more awareness, earlier diagnosis, and availability of better care in developed regions.

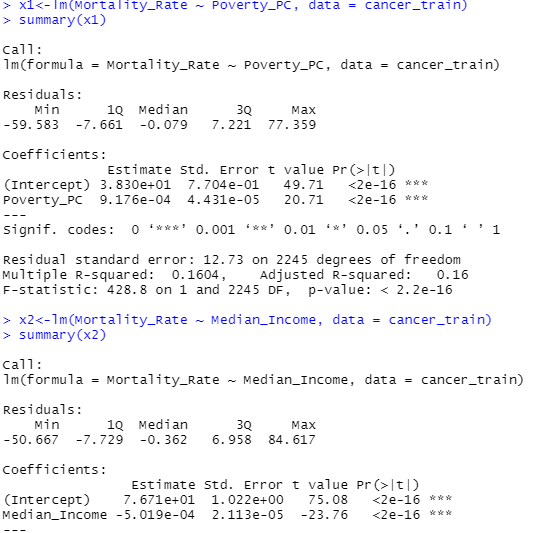
As expected, having no health insurance increases the mortality rate among the population whilst having it decreases the chances of death from cancer.

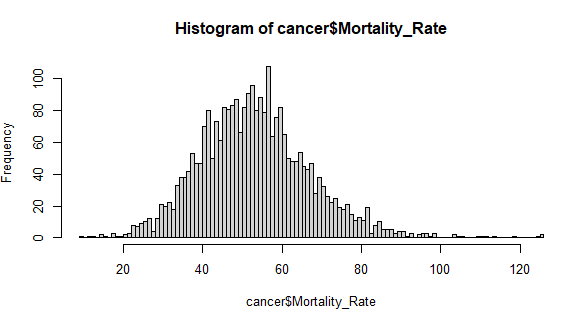
The Poverty Rate per capita also increases the chances of cancer fatality as it increases.



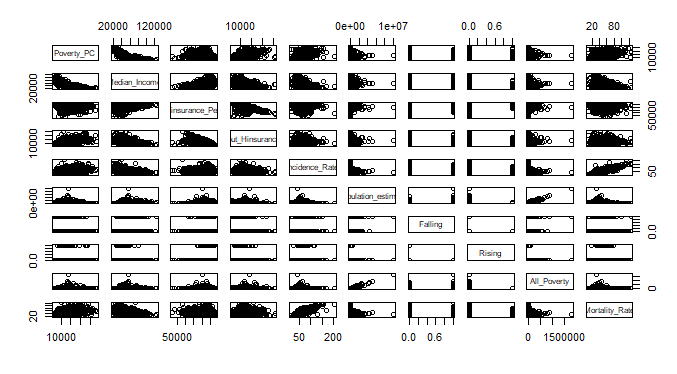
****



****



The Mortality Rate has a slightly right skewed shape so a transformation might be needed for this variable to reduce heteroscedasticity and improve the accuracy of the model.



The pairs plot enables us to visualize he correlations between the variables. Further below, we might remove the “Poverty” variable.

**Study Model:**

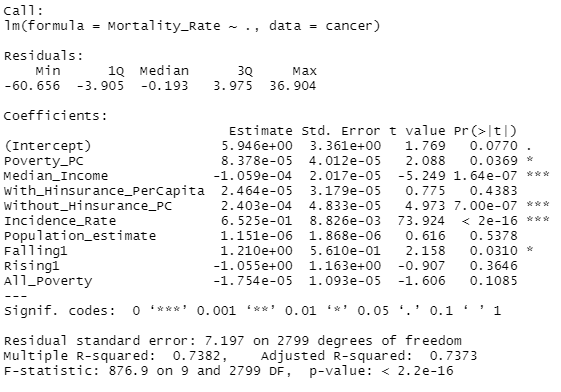
Based on the literature review, it has been the case that many social and economic features may impact a country’s reaction and treatment of cancer diagnoses, in this study we are focusing on eight of those factors.

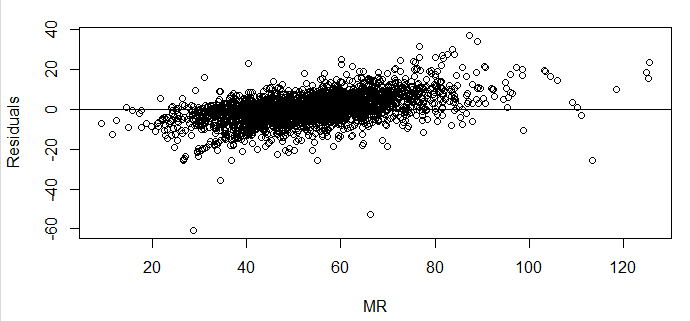
**Model A:**

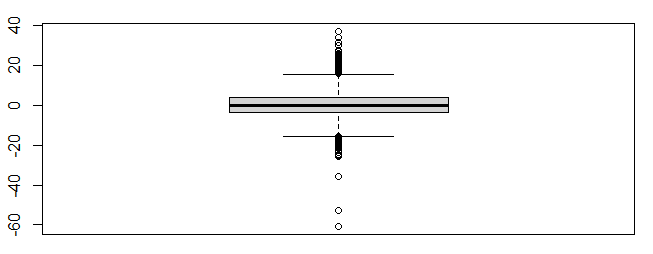
MR = β0 +β1P+ β2Pop + β3HI + β4noHI + β5MI + β6IR + β7R + β8F+ β9PPC

MR= Mortality Rate, P= Poverty Rate, Pop= Population Estimate, HI= people with health insurance per capita, noHI = people without health insurance per capita, MI= the median income, IR = Incidence Rate, R = Rising, F = Falling, PPC=Poverty Rate Per Capita.

In order to assess whether the independent variables are significant we need to look at several statistics including the R square and p values and certain assumptions must be made to test the model. In this case some assumptions are violated. It appears that there is no relation between some variables and the mortality rate which suggests that this model is not the best fit. For the equality of variances condition there is a tendency for the model to underestimate higher mortality rates.







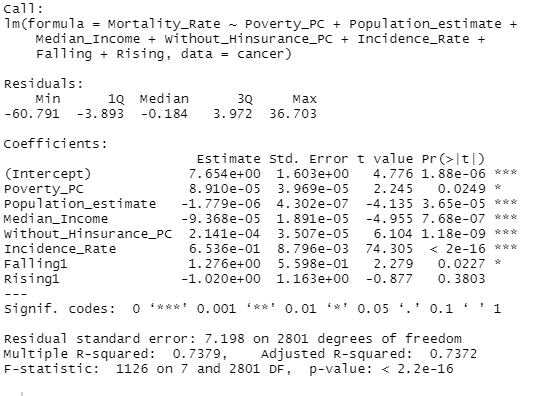
The Residuals vs Fitted graph above indicate clearly the presence of some outliers as well as the QQnorm, and the Cook’s distance indicated some influential points close to 0.4. The AIC for this model is 19071.29.

**Model B:**

MR = β0 + β1Pop + β2noHI + β3MI + β4IR + β5R + β6F+ β7PPC

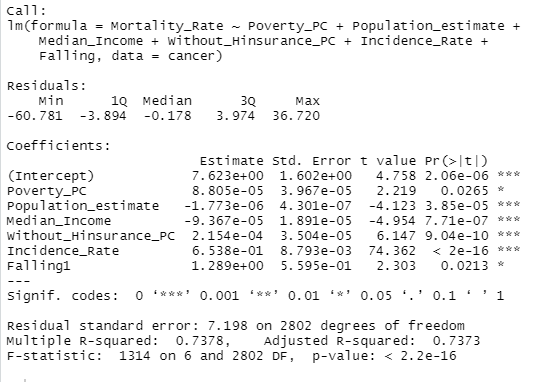
After examining the VIF score for model A, removing the variable “All\_Poverty” (VIF=23) is logical since it is likely correlated the rest of the variables.

After iteratively removing variables with high VIF scores and computing the results of the combinations, it was decided to remove the “number of people with health insurance per capita” variable which yielded a model with more significant results and with a VIF score lower than 5 for all variables.



This model’s parameters seem more logical, for example a higher median income means a lower mortality rate, and the more people without health insurance the larger the mortality rate.

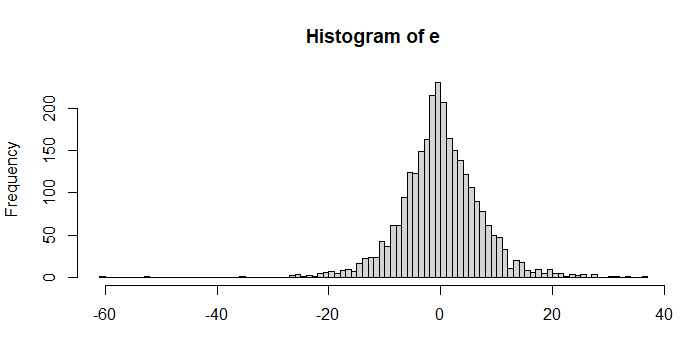
And Removing the “Rising Variable” yields this result:

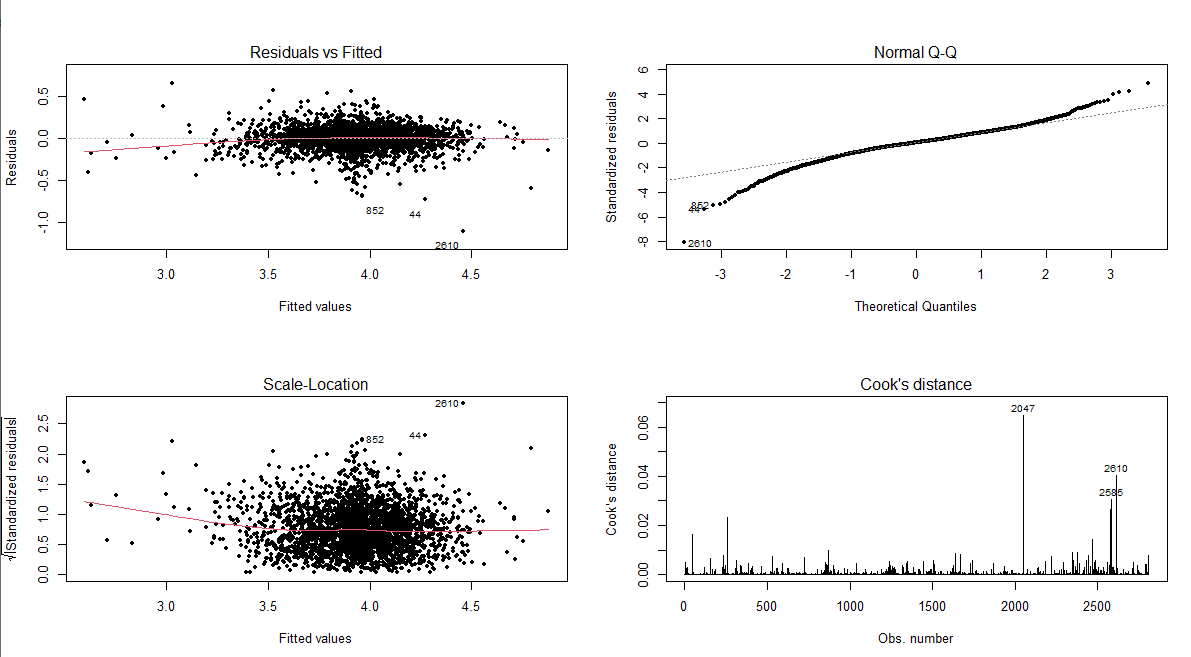


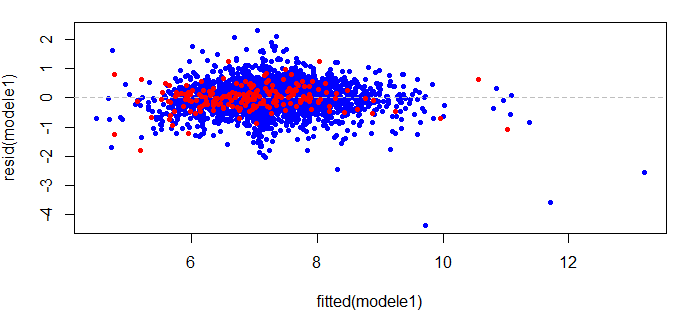
This summary shows that poverty per capita, median income, incidence rate, and no health insurance have a strong significance on predicting the mortality rate, however a recent trend of falling cases and the poverty per capita do not have strong influence. The R-square = 73.78% of the variance in the mortality rate was explained by model B, AIC: 19069.21, and p< 0.001. Furthermore, using the stepwise backward function yields the same model.

**Residuals Analysis:**

The distribution of the errors is not completely normal as indicated by the Shapiro-Wilk test: p= 2.2e-16. Visually, there is a small number of outliers to the left and the tails appear slightly wider than a normal distribution. Furthermore, the autocorrelation test indicated some correlation between the residuals.



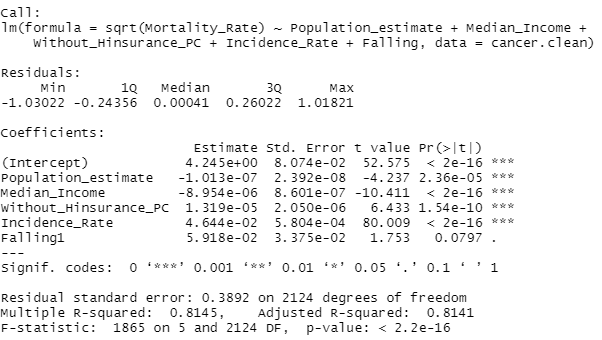




There is no visible difference in the residual’s vs fitted plot for Falling=”1” and Falling = “0”. Both models tend to underestimate in areas of high mortality rate.

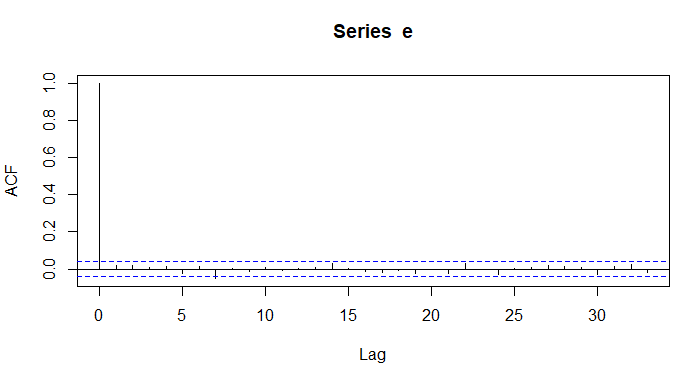
**Model C:**

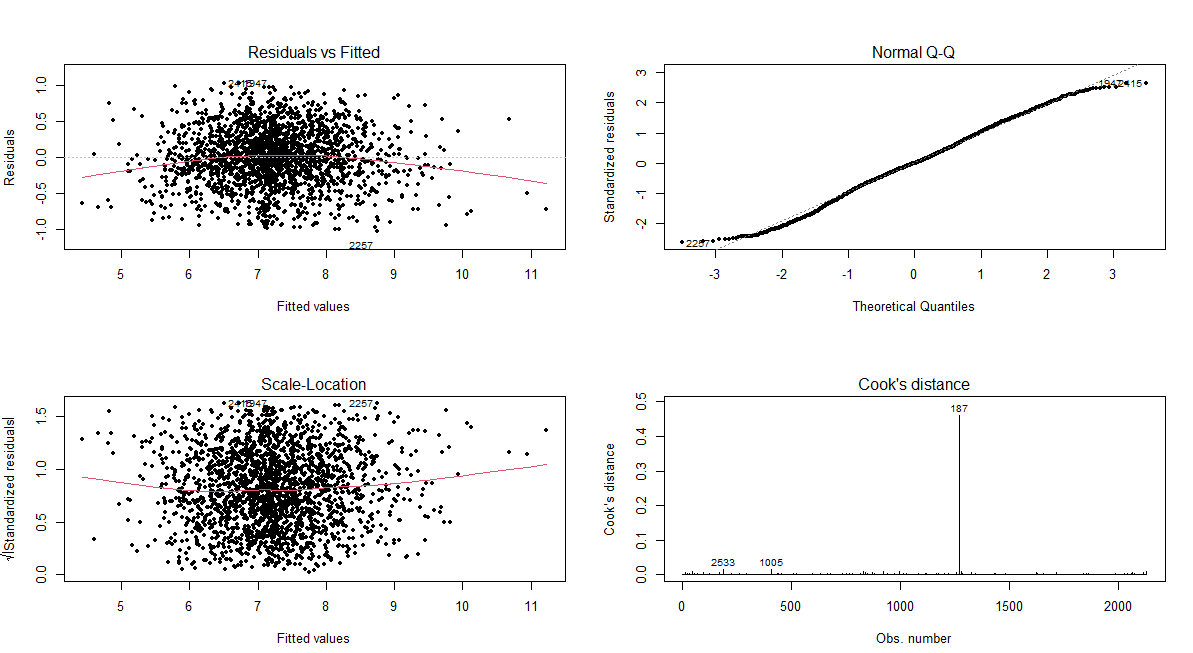
Sqrt(MR) = β0 + β1Pop + β2noHI + β3MI + β4IR + β5R + β6F



R square = 81% of the variance in mortality rate is explained by the model, p<0.001, and AIC=2033.

Using sqrt(y) instead of y helps to reduce the helps to reduce the heteroscedasticity of the residuals. However, the variable poverty per capita that was previously weakly significant had a p-value above 10% and had to be removed.





After removing the points having a studentized residual above the threshold 1.9 and below -1.9 the new model has normal residual values with no autocorrelation.

The AIC of this model is 2033, so model C, having the lowest AIC and the highest R square will be considered the best fitted model. 84% of the actual values in the testing set were found to be within the confidence interval. The RMSE = 0.38

**Conclusion:**

Based on the results of regression, five determinants: Median Income, Population Estimate, falling trend, Health Insurance, and Incidence Rate have a significant impact on Mortality Rate. This emphasizes the importance of raising awareness in underprivileged counties about the importance of good health habits and the use of healthcare to decrease the economic inequality of incidence and mortality rates in cancer.

**References**:

Income and Poverty in the United States: 2015

<https://www.census.gov/library/publications/2016/demo/p60-256.html>

10 Open Datasets For Linear Regression

TELUS International

<https://www.telusinternational.com/articles/10-open-datasets-for-linear-regression>

Factors Associated With Cancer Disparities Among Low-, Medium-, and High-Income US Counties

O'Connor et al.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6324449/>

Socioeconomic inequalities in lung cancer - a time trend analysis with German health insurance data

Tetzlaff F;Epping J;Tetzlaff J;Golpon H;Geyer S;

https://pubmed.ncbi.nlm.nih.gov/33740928/

Factors Linking Disparities in Cancer Deaths, Income

September 28 et al.

https://www.cancer.gov/news-events/cancer-currents-blog/2018/factors-linking-cancer-death-income-disparities