**Analysis of Cancer Mortality Rates for US**

**Team：Aces**

**Team Member: Jieling Ren, Xin Wen**

**Catalog**

[Data Overview](#_Toc600947735_WPSOffice_Level1) [3](#_Toc600947735_WPSOffice_Level1)

[Data Descriptive Statistics](#_Toc915402681_WPSOffice_Level1) [5](#_Toc915402681_WPSOffice_Level1)

[Classification](#_Toc600012459_WPSOffice_Level1) [12](#_Toc600012459_WPSOffice_Level1)

[Regression](#_Toc1973675748_WPSOffice_Level1) [19](#_Toc1973675748_WPSOffice_Level1)

# Data Overview

This data set contains multiple cancer trials data of US counties extracted from three separate sources, research information from clinicaltrials.gov, the incidence of cancer and death rates from cancer.gov, and economic and population statistics from census.gov. The original data cancer\_reg.csv comprises 34 Variables, 3047 Observations, records average numeric variables from 2010-2016. The demographics are from 2013 Census Estimates.

We are interested in the healthcare field, trying to analyze one or several typical diseases combining diverse factors. Besides, R is a good research tool for the medical and health industry. Therefore, we decided to choose this multi-dimensional cancer dataset to apply the statistical methods and machine learning we have learned to discover the theory and reason the resultant patterns out.

We gained this data from data.world, and the raw data are all from .gov websites which are authoritative and reliable resources. We also checked the accuracy of the original file on these government websites.

The original file is from https://data.world/nrippner/ols-regression-challenge.

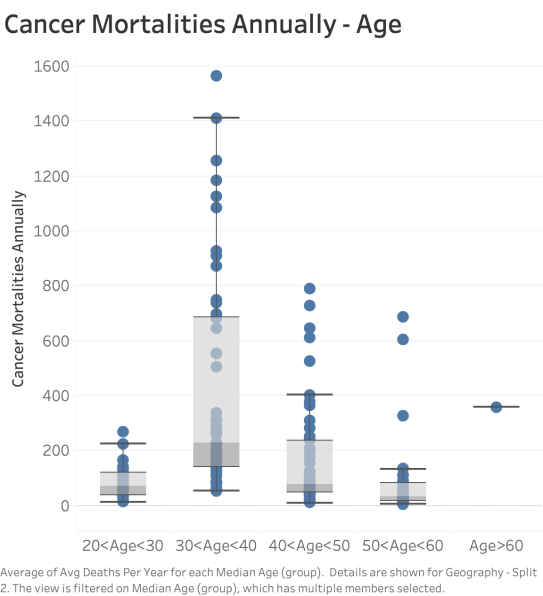
Table 1. Data Schema

|  |  |
| --- | --- |
| **Table Name** | **Explanation** |
| TARGET\_deathRate | Dependent variable. Mean per capita (100,000) cancer mortalities(a) |
| avgAnnCount | Mean number of reported cases of cancer diagnosed annually(a) |
| avgDeathsPerYear | Mean number of reported mortalities due to cancer(a) |
| incidenceRate | Mean per capita (100,000) cancer diagoses(a) |
| medianIncome | Median income per county (b) |
| popEst2015 | Population of county (b) |
| povertyPercent | Percent of populace in poverty (b) |
| studyPerCap | Per capita number of cancer-related clinical trials per county (a) |
| binnedInc | Median income per capita binned by decile (b) |
| MedianAge | Median age of county residents (b) |
| MedianAgeMale | Median age of male county residents (b) |
| MedianAgeFemale | Median age of female county residents (b) |
| Geography | County name (b) |
| AvgHouseholdSize | Mean household size of county (b) |
| PercentMarried | Percent of county residents who are married (b) |
| PctNoHS18\_24 | Percent of county residents ages 18-24 highest education attained: less than high school (b) |
| PctHS18\_24 | Percent of county residents ages 18-24 highest education attained: high school diploma (b) |
| PctSomeCol18\_24 | Percent of county residents ages 18-24 highest education attained: some college (b) |
| PctBachDeg18\_24 | Percent of county residents ages 18-24 highest education attained: bachelor's degree (b) |
| PctHS25\_Over | Percent of county residents ages 25 and over highest education attained: high school diploma (b) |
| PctBachDeg25\_Over | Percent of county residents ages 25 and over highest education attained: bachelor's degree (b) |
| PctEmployed16\_Over | Percent of county residents ages 16 and over employed (b) |
| PctUnemployed16\_Over | Percent of county residents ages 16 and over unemployed (b) |
| PctPrivateCoverage | Percent of county residents with private health coverage (b) |
| PctPrivateCoverageAlone | Percent of county residents with private health coverage alone (no public assistance) (b) |
| PctEmpPrivCoverage | Percent of county residents with employee-provided private health coverage (b) |
| PctPublicCoverage | Percent of county residents with government-provided health coverage (b) |
| PctPubliceCoverageAlone | Percent of county residents with government-provided health coverage alone (b) |
| PctWhite | Percent of county residents who identify as White (b) |
| PctBlack | Percent of county residents who identify as Black (b) |
| PctAsian | Percent of county residents who identify as Asian (b) |
| PctOtherRace | Percent of county residents who identify in a category which is not White, Black, or Asian (b) |
| PctMarriedHouseholds | Percent of married households (b) |
| BirthRate | Number of live births relative to number of women in county (b) |
|  |  |

# Data Descriptive Statistics

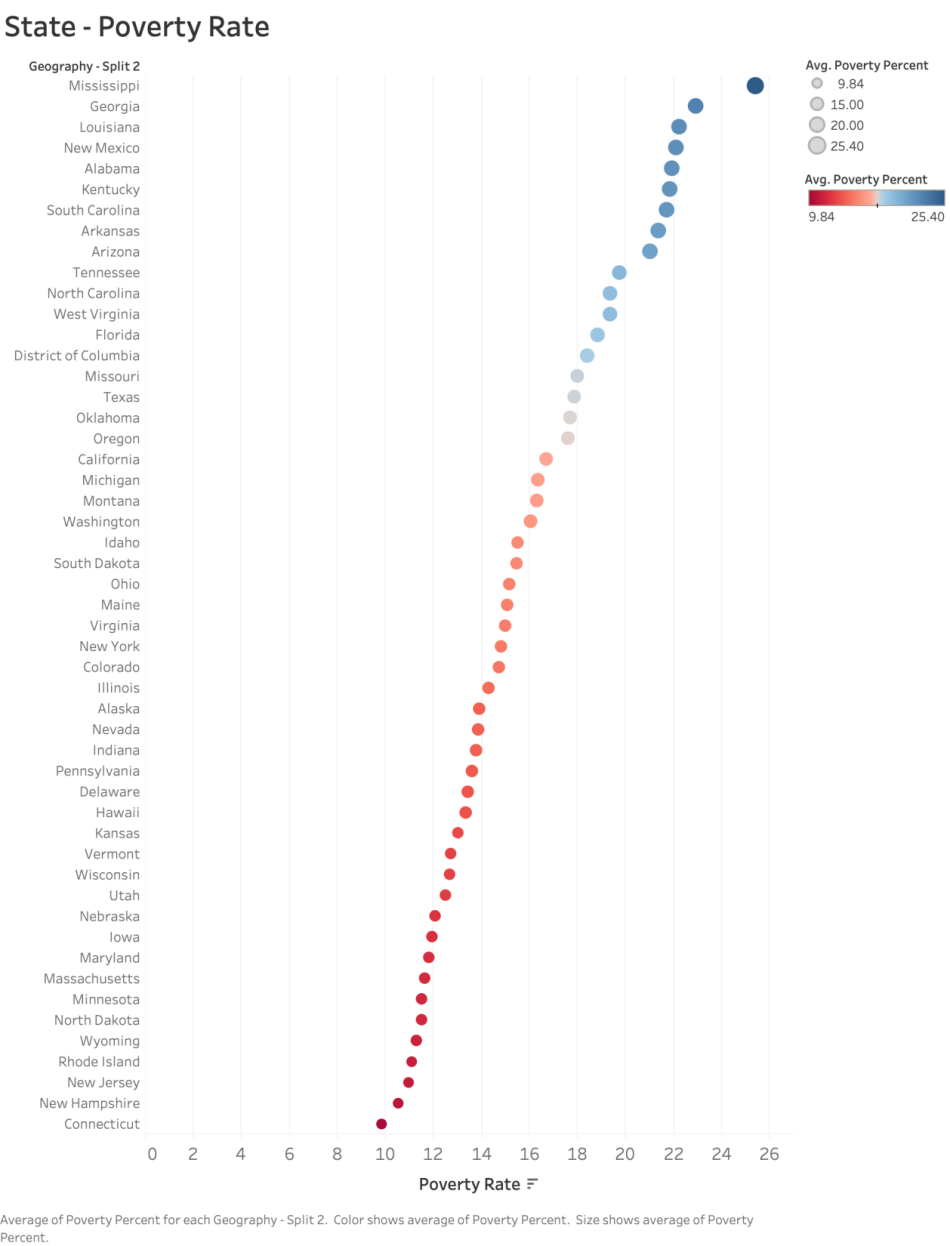
## Cancer Mortalities vs Age

We prepared the Age for grouping and deleted the oultiers. Since there is only one data for age above 60, it is relatively small, but we also keep it to reflect the full picture of the data. It can be clearly seen in the graph that in the age group 30-40, Cancer Mortalities is the highest, which is a really high related index of cancer like we expected before.



## State vs Poverty Rate

In addition to the direct factors related to cancer, we considered some macroscopic data of potential elements that could affect the incidence rate indirectly like poverty. The chart below shows the per capita poverty level in different states, which may reflect people's living conditions from the side. Poverty could have an influence on people’s health status and health insurance. People who earn less could take the most dangerous and harmful jobs which probably cause them body damage. And once they don’t have enough coverage and deposit to pay for the hospital for examination or treatment, the mortality of cancer could be very high.



## Statistical Description of Important Variables

The minimum, maximum, and average (mean, median, mode) and standard deviation of important variables are as follow.

The ***TARGET\_deathRate*** means the mean per capita (100,000) cancer mortalities(a). For the TARGET\_deathRate, the minimum is 59.7, the maximum is 362.8, the mean is 178.7, the median is 178.1, the mode is 184.3, the standard deviation is 27.75.

The ***incidenceRate*** means the mean per capita (100,000) cancer diagoses(a). For the incidenceRate, the minimum is 201.3, the maximum is 1206.9, the mean is 448.4, the median is 453.5, the mode is 453.55, the standard deviation is 54.56.

The ***medIncome*** means the median income per county (b). For the medIncome, the minimum is 22640, the maximum is 125635, the mean is 47063, the median is 45207, the mode is 34116, the standard deviation is 12040.09.

The ***PctPrivateCoverage*** means the percent of county residents with private health coverage (b). For the PctPrivateCoverage, the minimum is 22.30, the maximum is 92.30, the mean is 64.35, the median is 65.10, the mode is 65.30, the standard deviation is 10.65.

The ***PctPubliceCoverageAlone*** means the percent of county residents with government-provided health coverage alone (b). For the PctPubliceCoverageAlone, the minimum is 2.60, the maximum is 46.60, the mean is 19.24, the median is 18.80, the mode is 18.70, the standard deviation is 6.11.

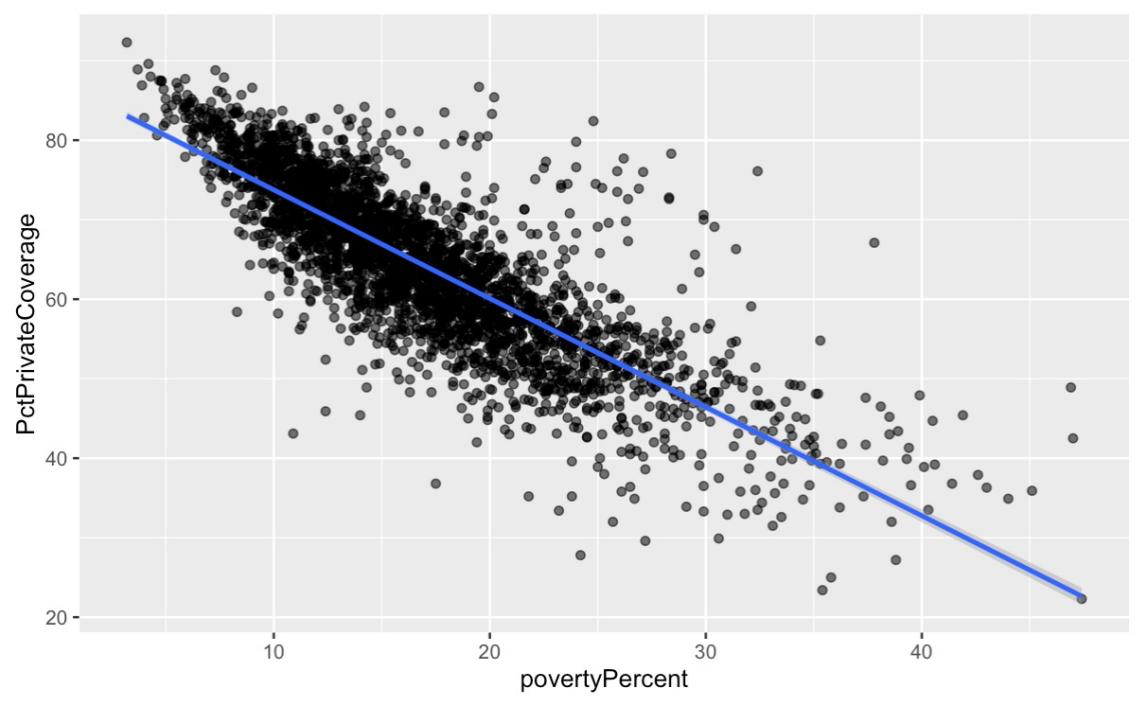
The ***povertyPercent*** means the percent of populace in poverty (b). For the povertyPercent, the minimum is 3.20, the maximum is 47.40, the mean is 16.88, the median is 15.90, the mode is 13.90, the standard deviation is 6.41.

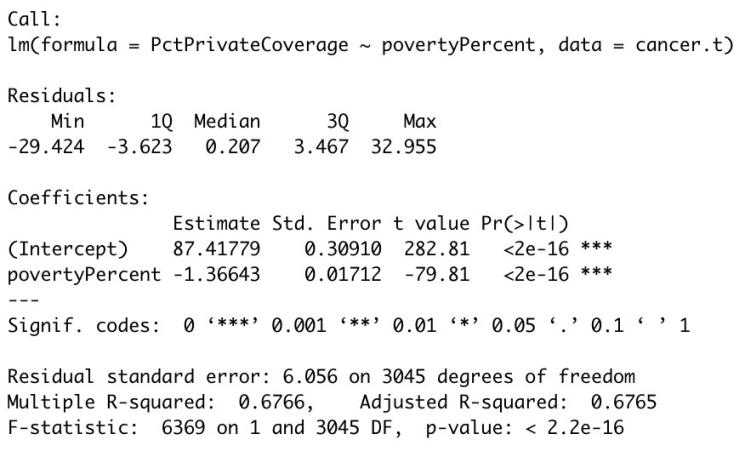
## Scatterplots

### Coverage vs Poverty Rate

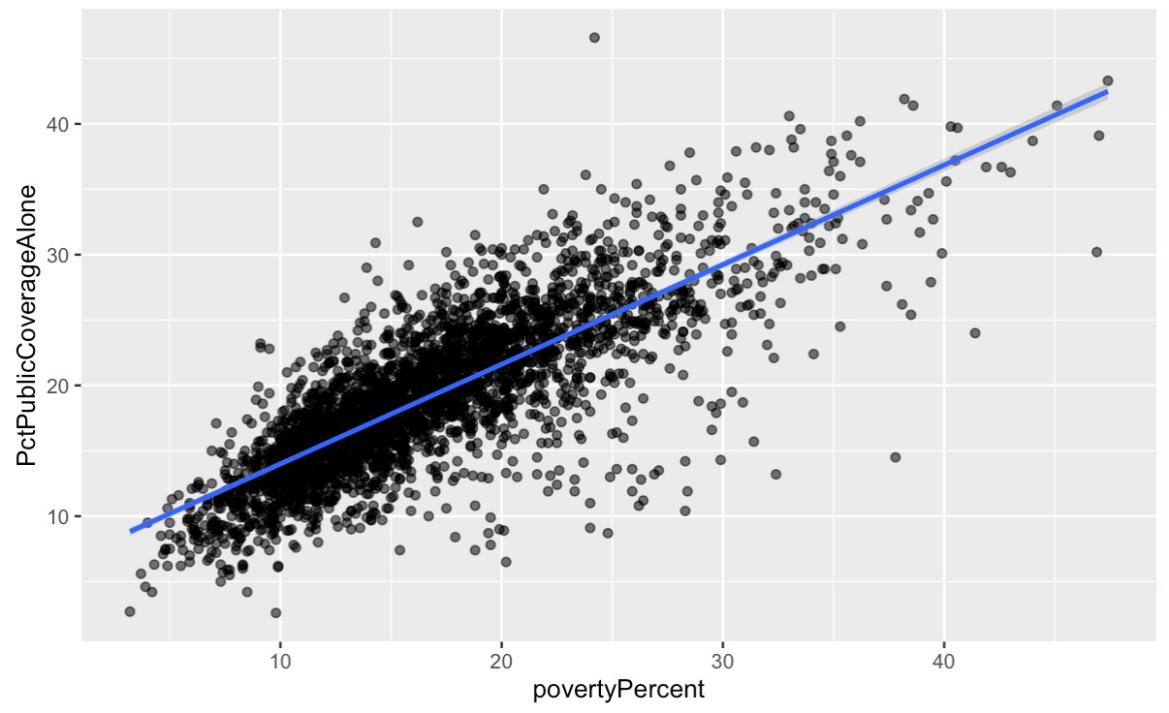
Most cancer patients will use medical insurance during treatment. Therefore, in these two graphs below, we consider the relationship between insurance and the poverty rate. It can be seen that the higher the poverty rate, the more common public insurance; the lower the poverty rate, the more common private insurance.

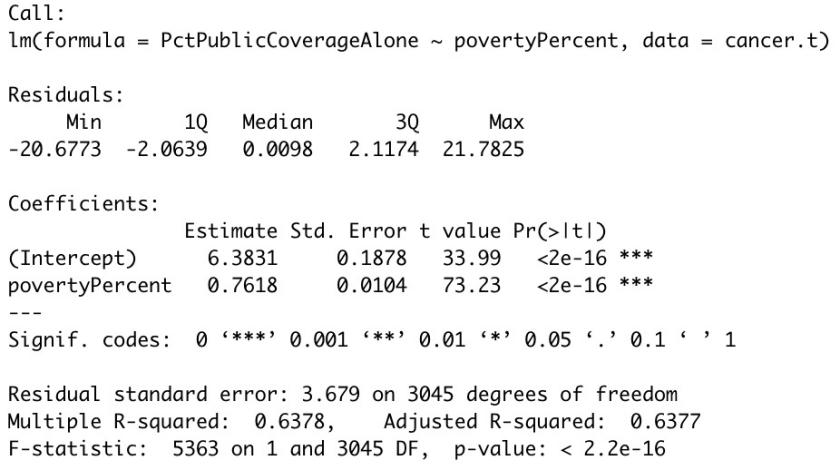
For the regression of Percentage of Private Coverage and Poverty Rate, the R- squared is 0.6766 and the adjusted R-squared is 0.6765. It means that the Percentage of Private Coverage can be 67.66% explained by the Poverty Rate. Also, the Poverty Rate is statistically significant as its p-value are far less than 0.05. We can see the coefficient of Poverty Rate is -1.36643 which means when the Poverty Rate increases one unit, the Percentage of Private Coverage will decrease 1.36643 units.





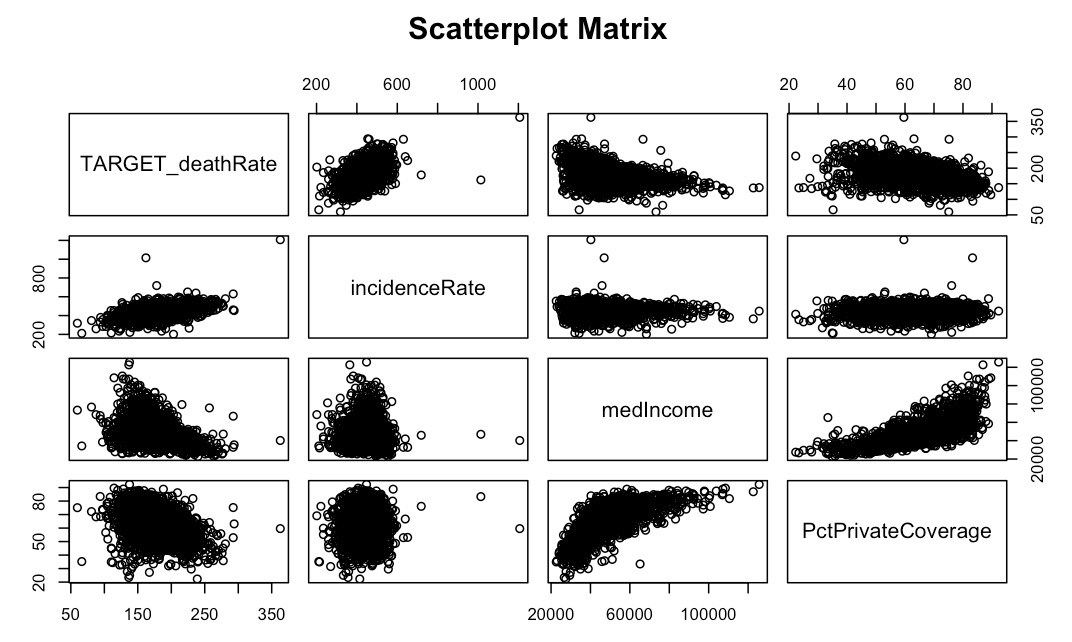
For the regression of Percentage of Public Coverage Alone and Poverty Rate, the R- squared is 0.6378 and the adjusted R-squared is 0.6377. It means that the Percentage of Public Coverage Alone can be 63.77% explained by the Poverty Rate. Also, the Poverty Rate is statistically significant as its p-value are far less than 0.05. We can see the coefficient of Poverty Rate is 0.7618 which means when the Poverty Rate increases one unit, the Percentage of Private Coverage will increase 0.7618 units.

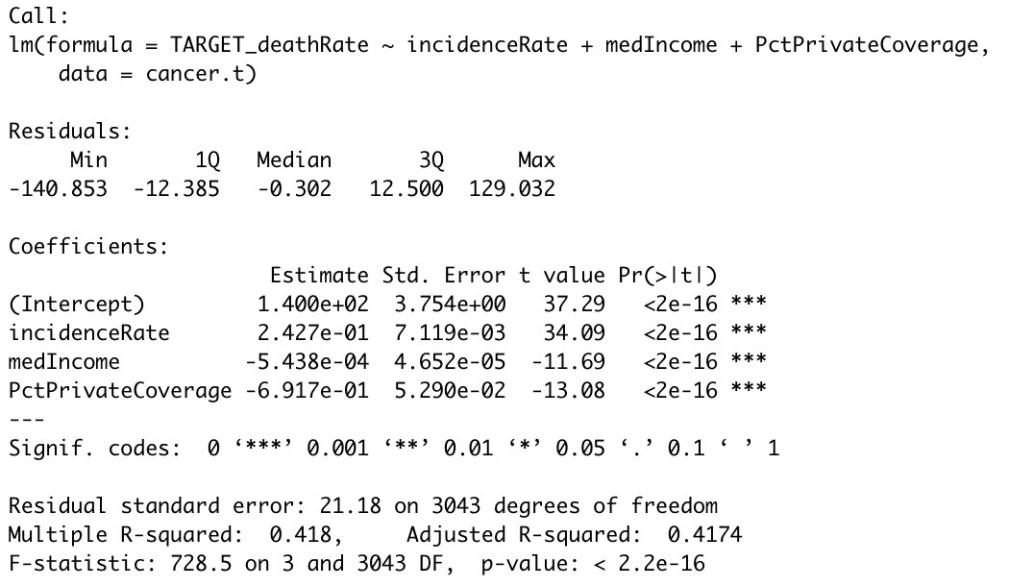


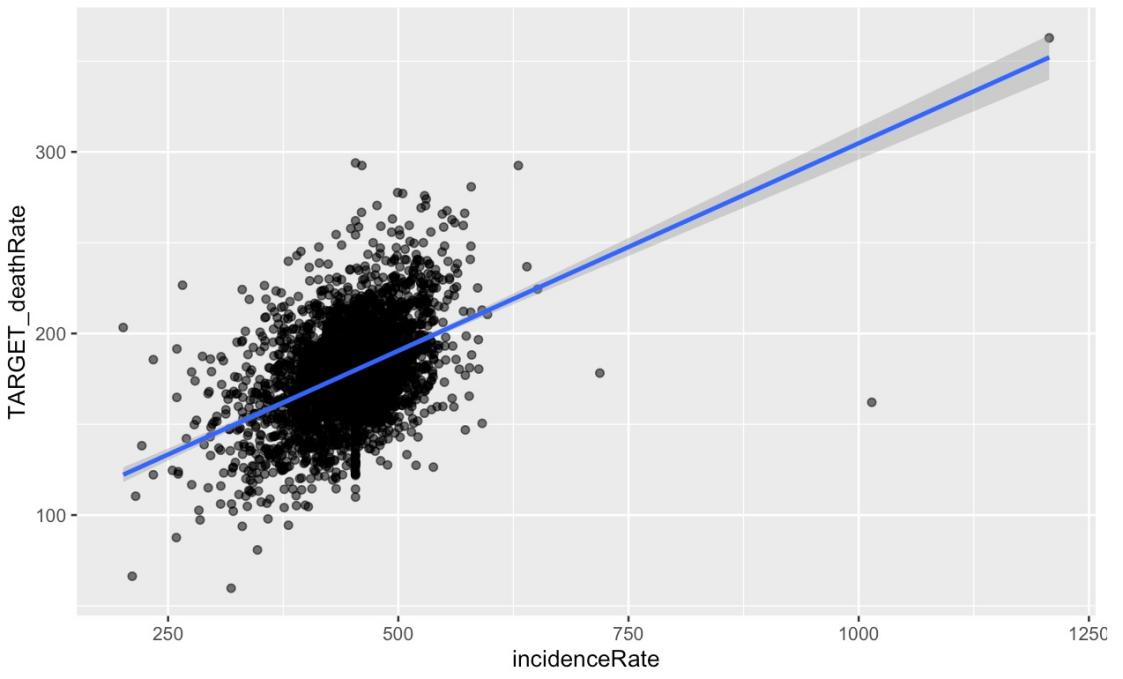


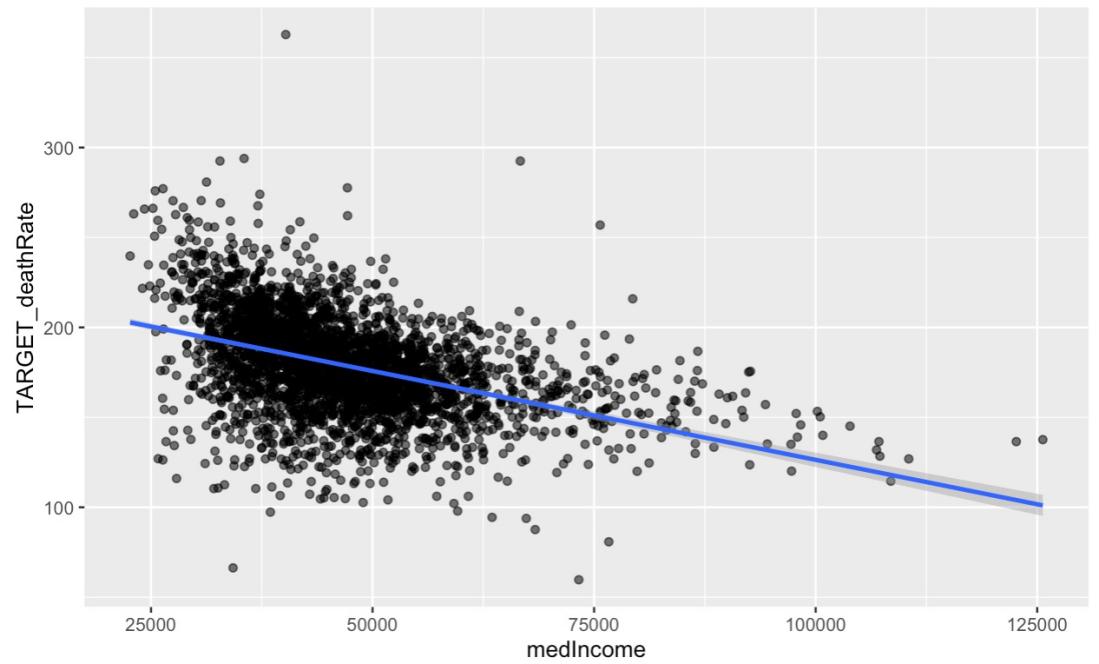
### Death Rate vs Incidence Rate, Median Income, Private Coverage

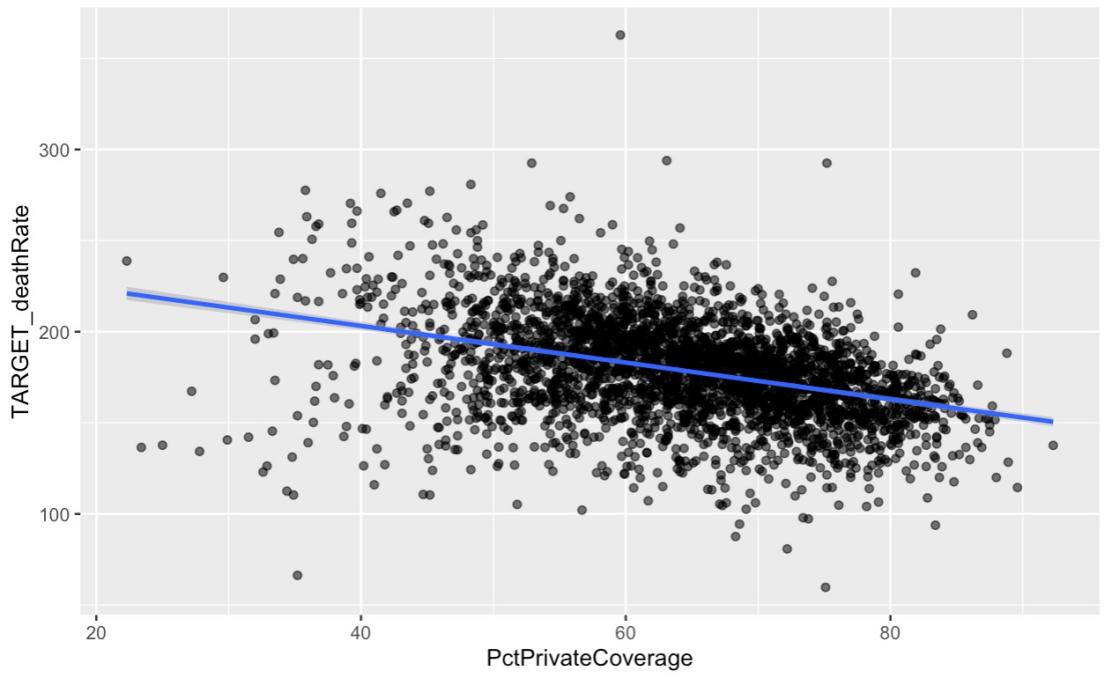
The regression result of Death Rate between Incidence Rate, Median Income, and Private Coverage is as follow. The R- squared is 0.418 and the adjusted R-squared is 0.4174. It means that the Death Rate can be 41.8% explained by the three independent variables. Also, the 3 independent variables are statistically significant as their p-value are far less than 0.05. We can see the coefficient of Incidence Rate is 2.427e-1 which means when the Incidence Rate increases one unit, the Death Rate will increase 2.427e-1 unit. It makes sense as the higher Incidence Rate, the higher the Death Rate. The coefficient of Median Income is -5.438e-4 which means when the Median Income increases one unit, the Death Rate will decrease 5.438e-4 unit. The reason may be that the higher income, the more money can be used to cure cancer. Thus, the Death Rate will decline with the higher Median Income. However, the influence of Median Income is smaller than the other two variables. The coefficient of Private Coverage is -6.917e-1 which means when the Private Coverage increases one unit, the Death Rate will decrease 6.917e-1 unit. This may be because the more Private Coverage the people buy, the more concerned they will be about their health. Thus, the Death Rate will decline with the higher Private Coverage.











# Classification

## Logistic Regression

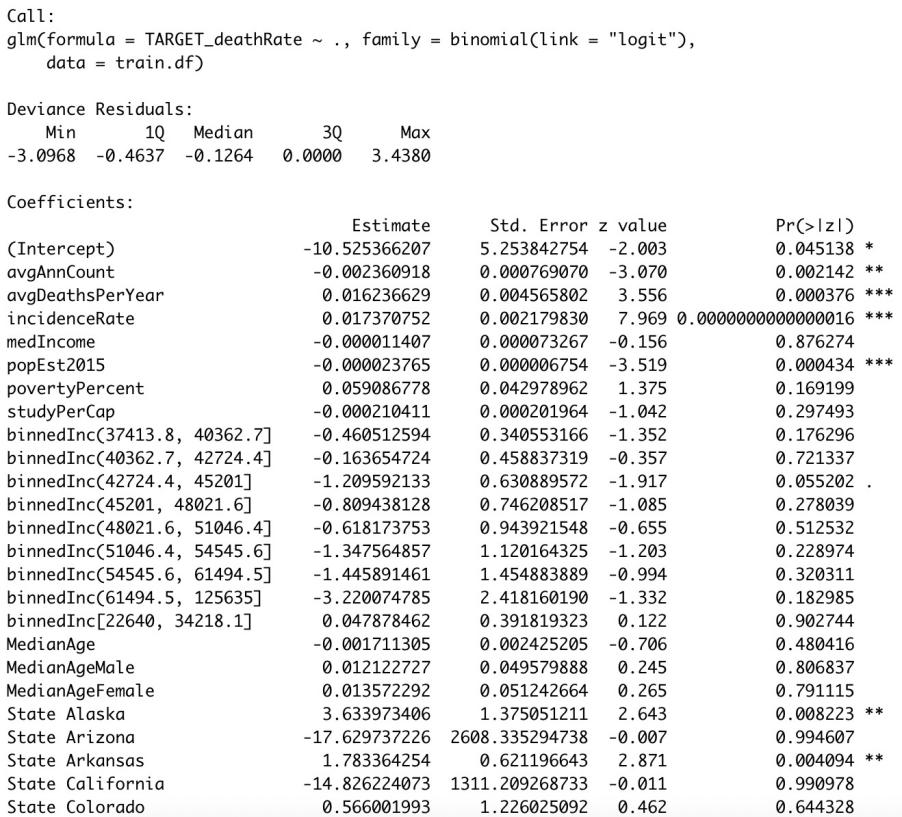
In this case, we select the ***TARGET\_deathRate*** as the target categorical variable. In raw material, the TARGET\_deathRate is a numeric variable that represents the mean cancer mortality per 100,000 people. We transform it into a binary target variable that includes 1 and 0 which means high mortality that is higher than 200 (the US average cancer mortality in 2015) and low mortality that is lower than 200 respectively.

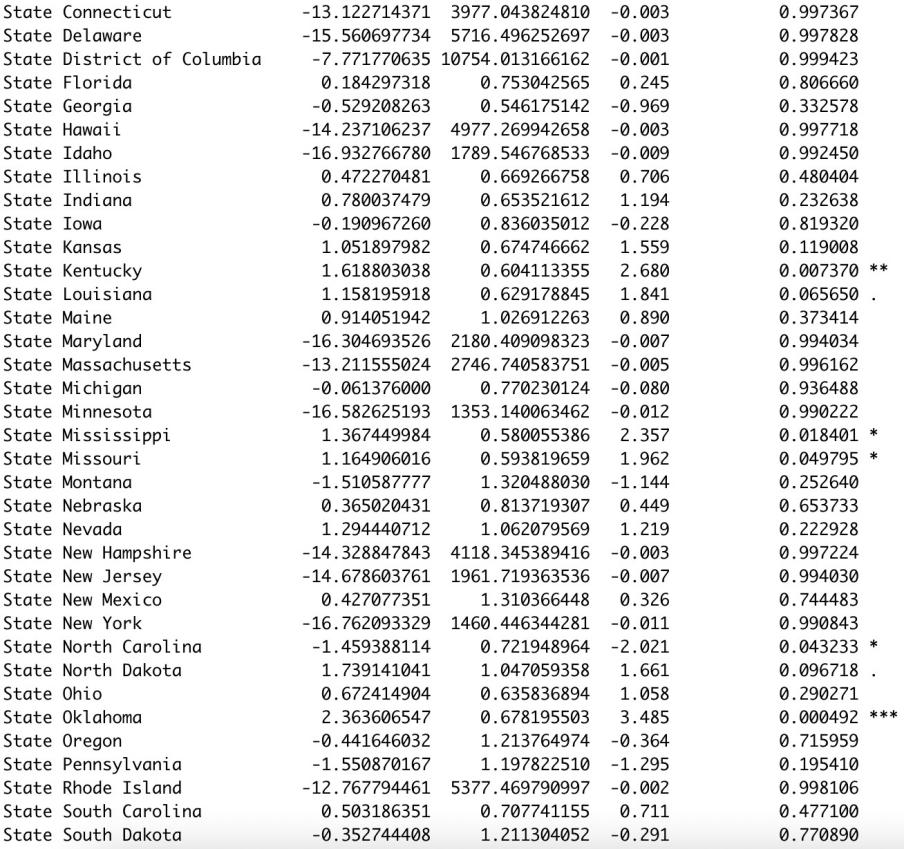
As we are interested in cancer mortality related to socioeconomic factors, the accurate classification of this target variable matters. Actually, TARGET\_deathRate was a numeric variable that might be more suit for linear regression, but we need an index to measure whether it is an area with a high risk of cancer mortality. So we can use this transformed variable to ignore an area with its degree of cancer mortality.

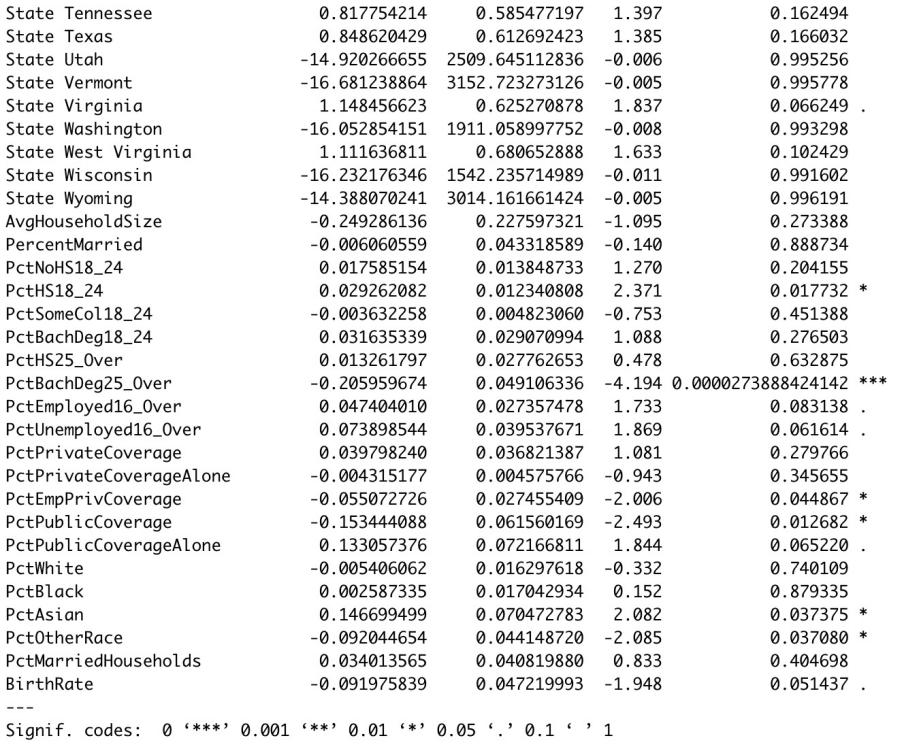
Before conducting logistic regression, we change binnedInc into a factor by different intervals and split the Geography column into two separate columns in order to figure out the relationship between state and cancer mortality.

Then, we drop the variables that are meaningless: "County/City".

We start with all the relevant variables.

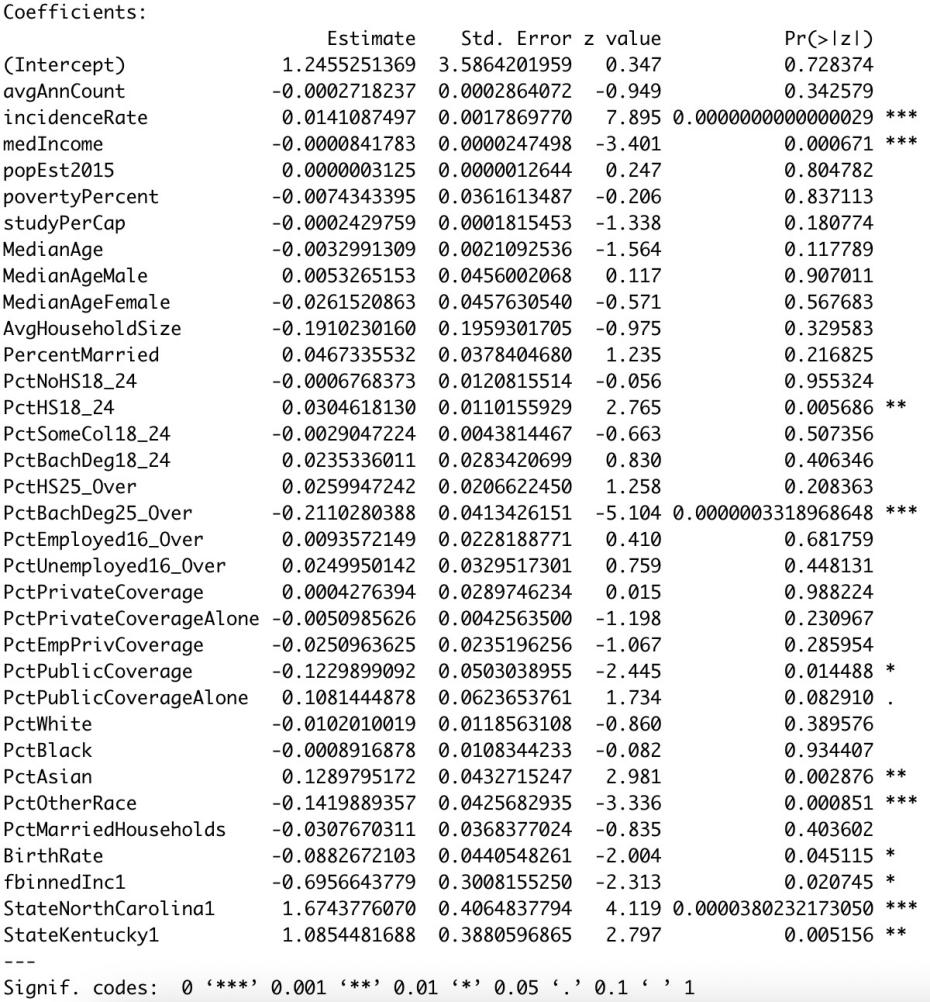






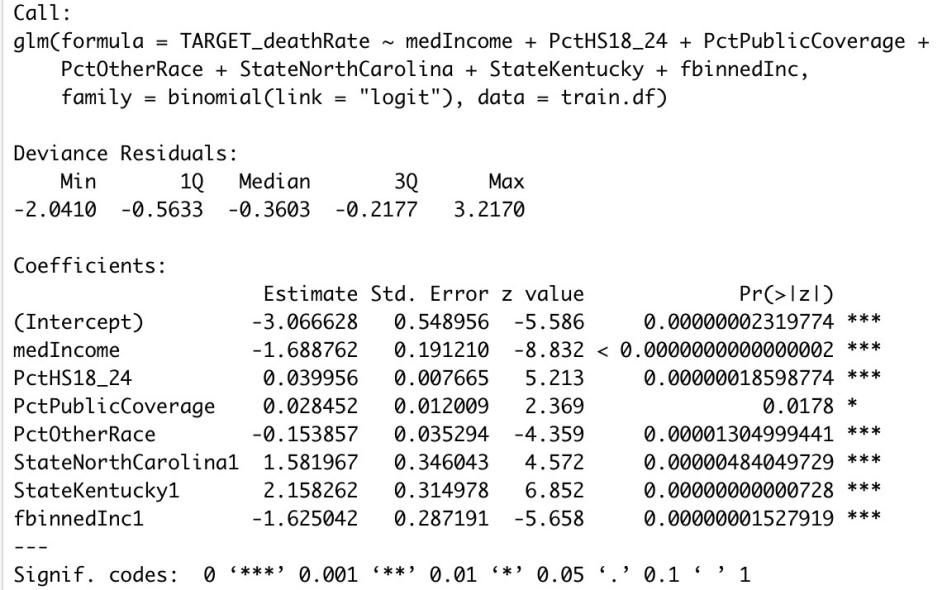
In the output above, according to the P-value of levels of binnedInc and State, we chose binnedInc == '(42724.4, 45201]', Oklahoma and Kentucky to create dummy variables staying in the algorithm. We change medIncome into a binary variable for high median income and low median income. And we drop the State, binnedInc, and the avgDeathsPerYear that are too good to fit.

Then we get the result below.

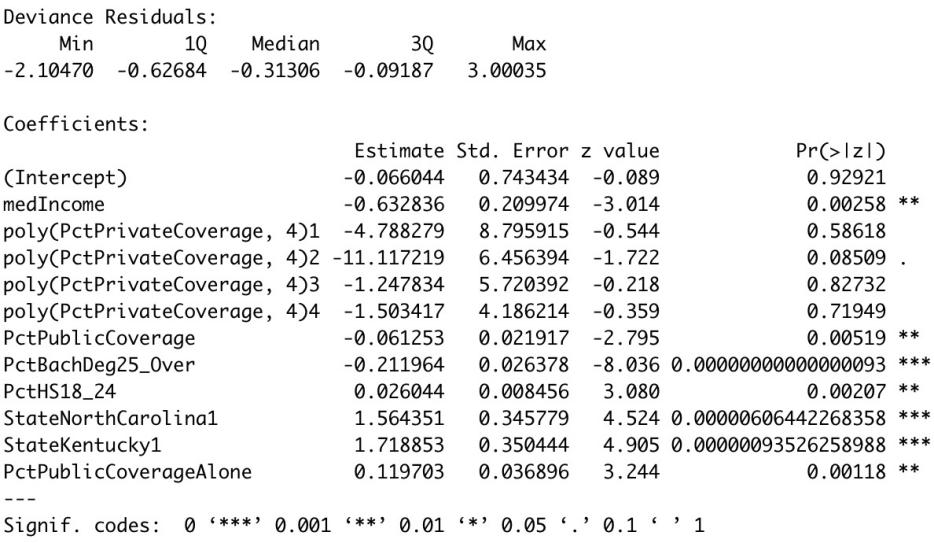


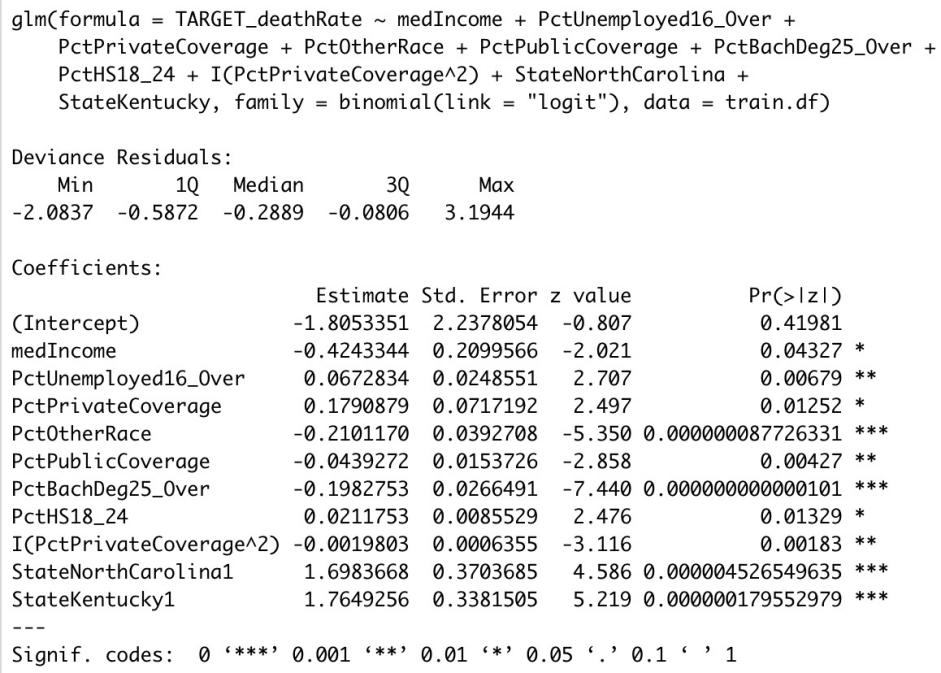
Continue to remove the variables that are too good to fit.

Now we create logistic model with only relevant variables.



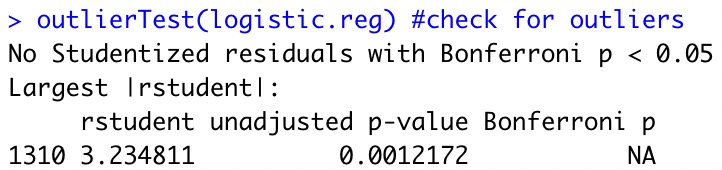
The result looks good, all the variables are statistically significant. We tried to find the interaction between variables, but all the potential interactions we’ve included all get a high P-value(>0.05), so we decided to add polynomial terms.

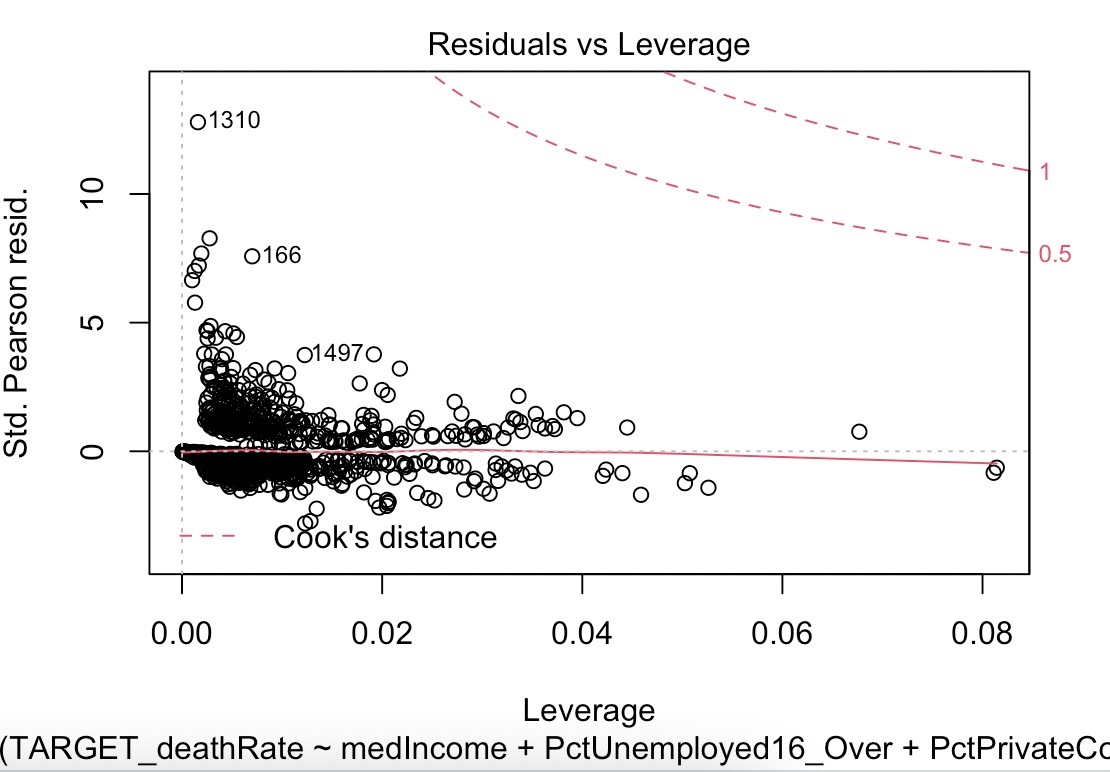


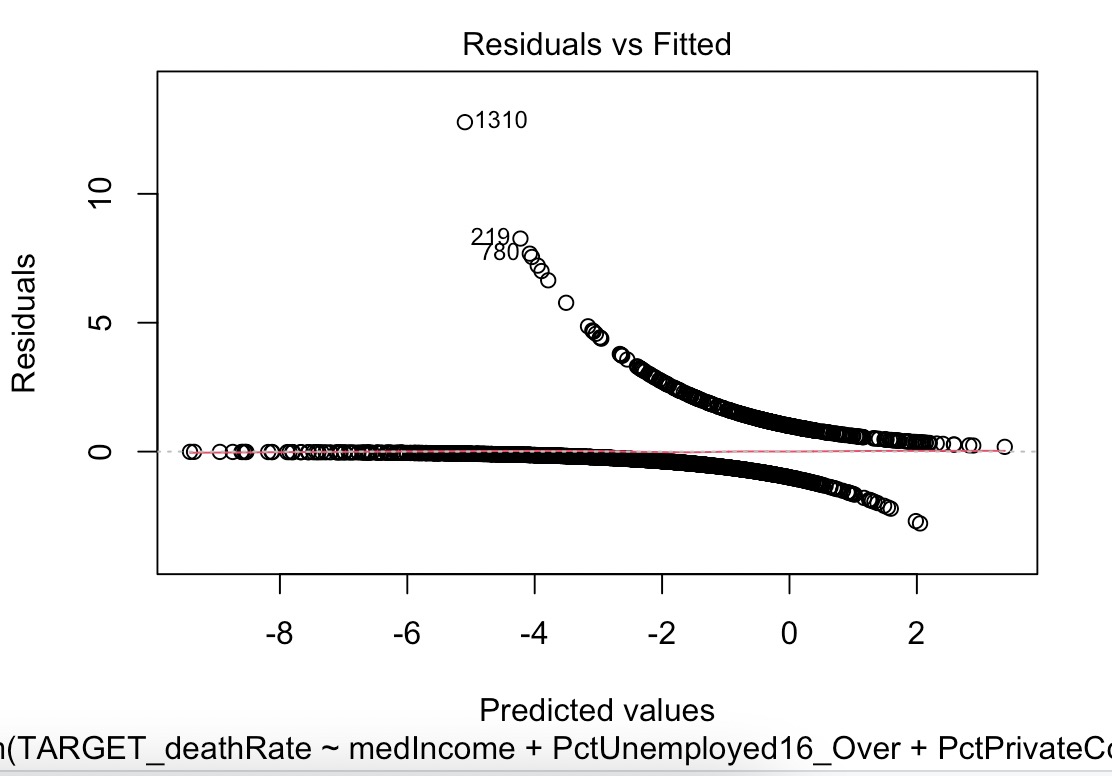


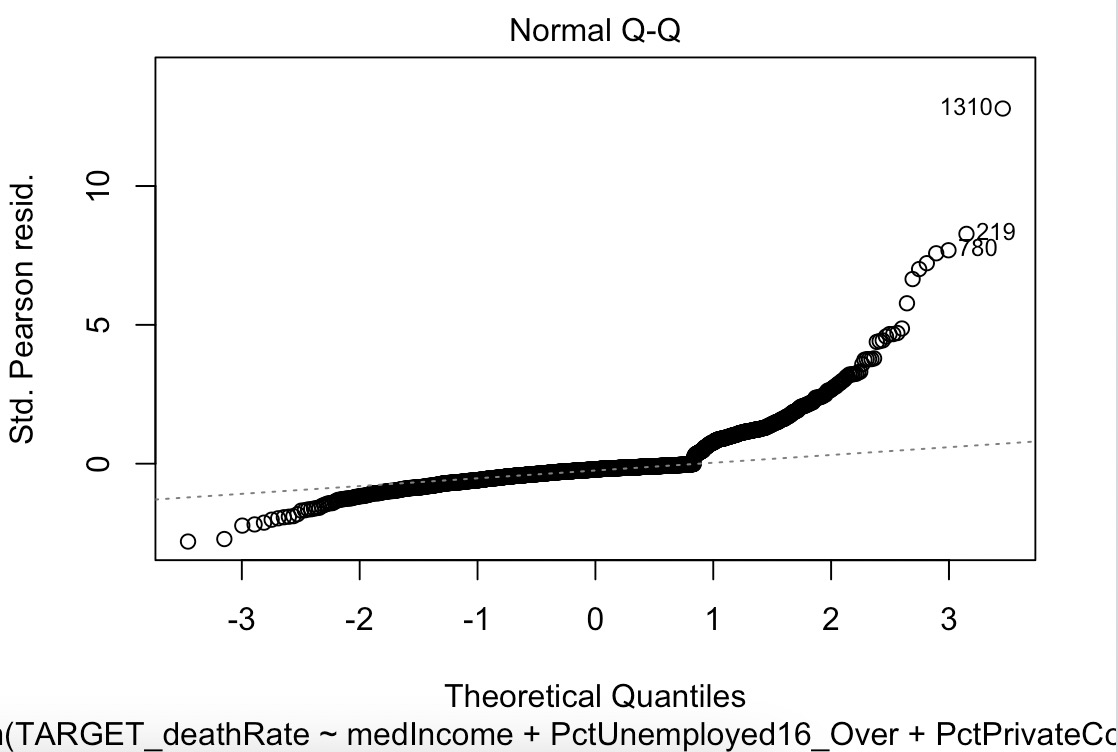
After using poly() function to check the polynomial from 2-4, the square term shows the best outcome. So we include the square of PctPrivateCoverage in the logistic regression.

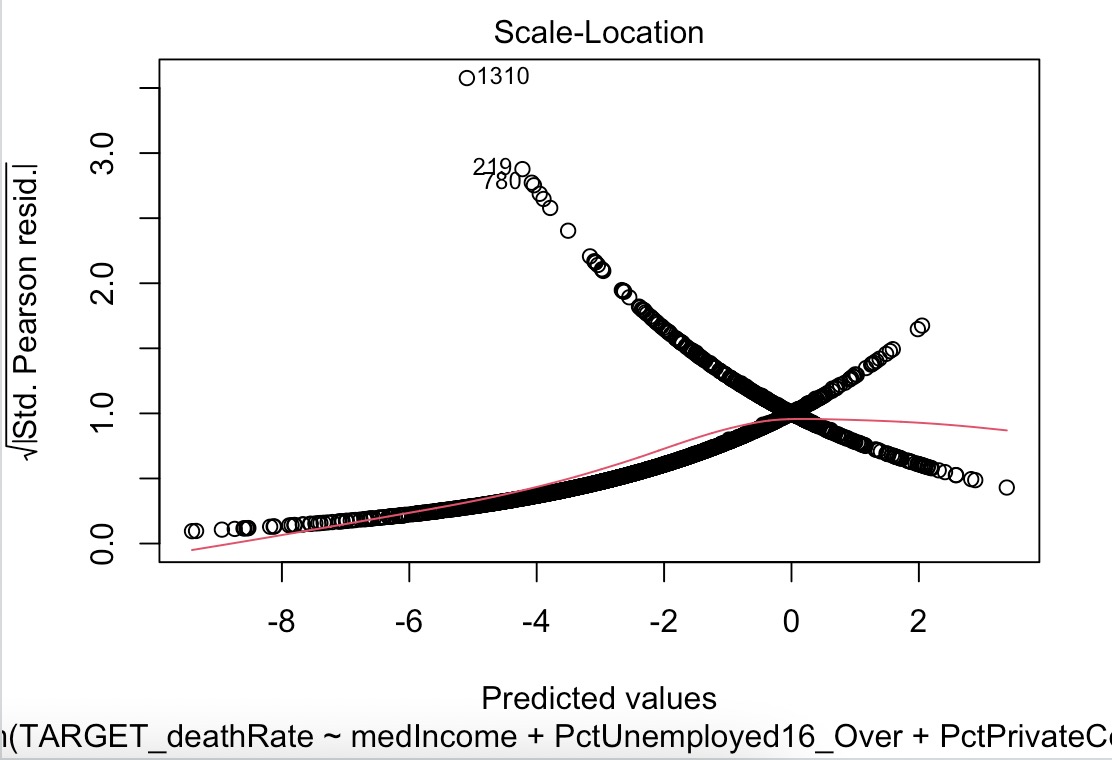
Check for outliers. Only for one 1310, the residuals align well, we can omit it.



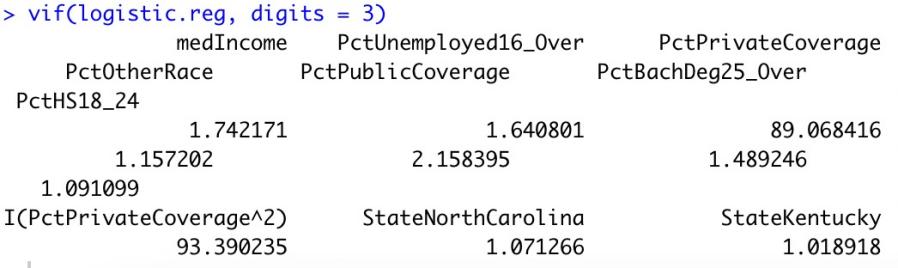




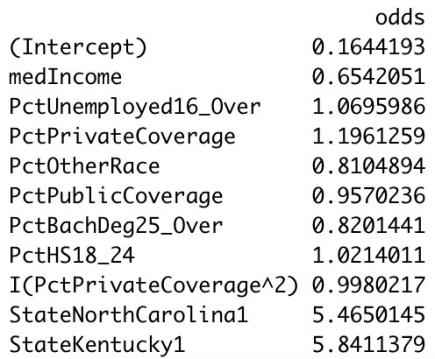




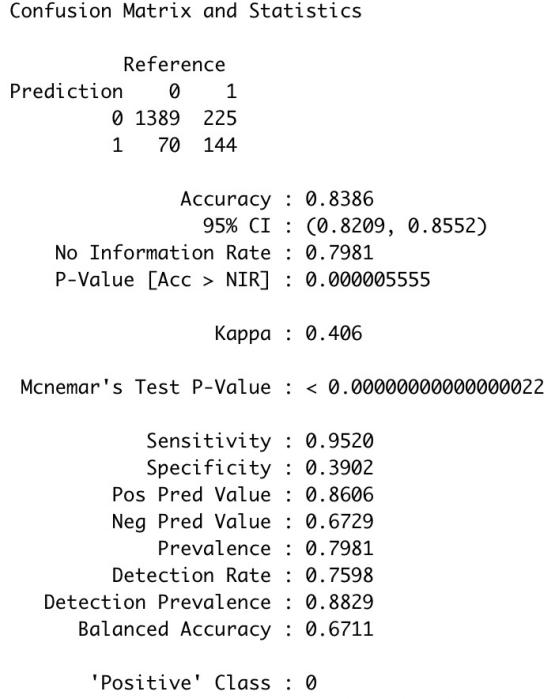
Check for col-linearity.



Interpretation by log odds:



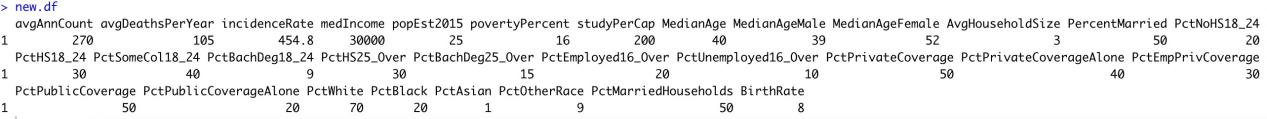
In this logistic regression, high median income decreases the probability of cancer mortality by 35%, maybe because more residents have money for treatment; to not being employed residents are likely to die for cancer by increasing probability of 0.07%, maybe the same reason as above. Owning only private coverage increases the probability of cancer mortality by 0.2 percent. Having more residents who are of other races may decrease the mortality by 0.2%. Residents who are over and gained a bachelor's degree decrease the probability by 0.18%. The highly educated people may know how to keep healthy better. In contrast, people who have a high school diploma increase the odds by0.02%. N.C and Kentucky will increase cancer mortality by over 500%.

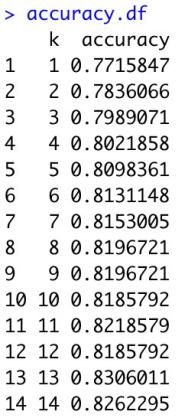


Through the confusion matrix and the statistic analysis, we see our model is not overfitting.

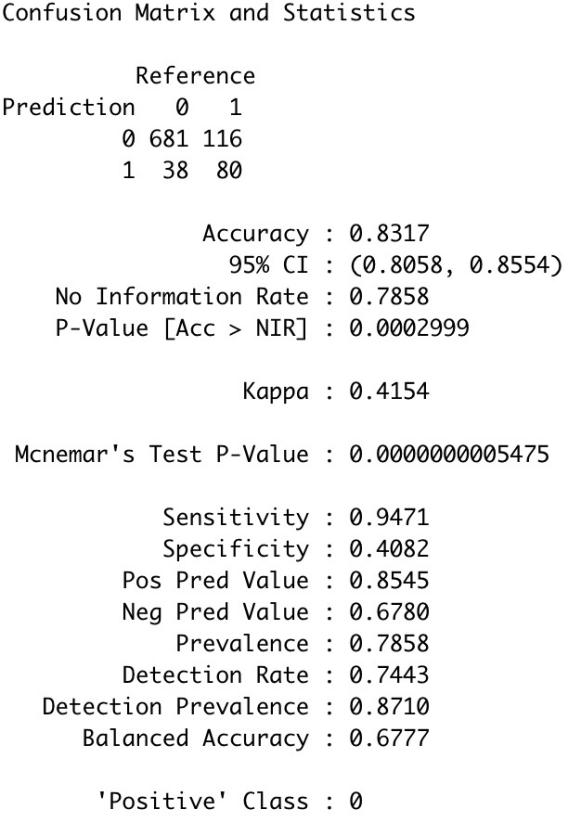
## KNN

For KNN classification, we include variables below.





Through the knn accuracy analysis, we found the best k is 13. And we got the confusion matrix and statistics analysis with this k.



Comparing the confusion matrices of the best models from logistic regression and knn, we consider that knn is better because logistic regression and knn have similar accuracies, but knn has a lower sensitivity and a higher specificity so that has a higher balanced accuracy. Since we are studying the mortality of cancer, we don’t need high sensitivity, but if we consider cancer detection, we need high sensitivity, and care about false negatives, because we don’t want anyone to be misdiagnosed.

## Conclusion

From our models, we can draw a conclusion that the majority of areas have cancer mortality lower than the mean level around America in 2015. The high cancer mortality is highly related to median income, education level, employment, health insurance coverage, etc. Government should care about the Socioeconomic status, try to grow local economies, increase employment, fund education, fund medical research, increase health insurance funding, introduce policies to help cancer patients with diagnosis and treatment.

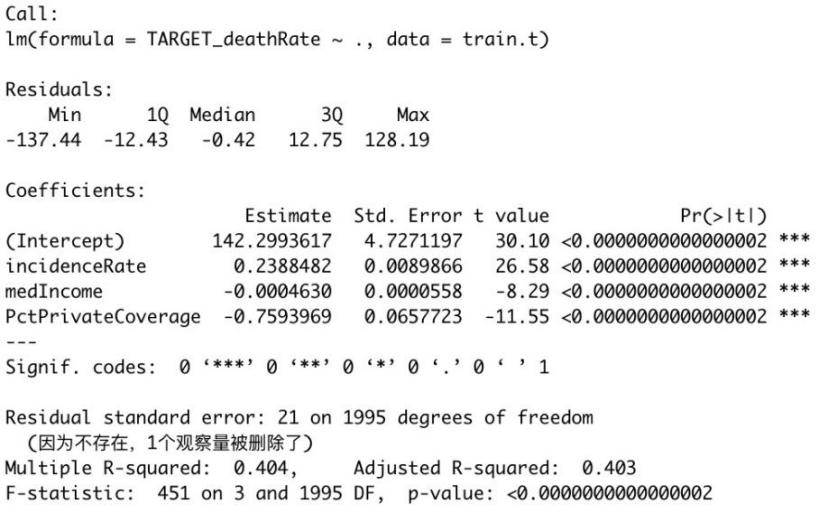
# Regression

We select the ***TARGET\_deathRate*** as our target categorical variable, which means the mean per capita (100,000) cancer mortalities(a). The goal of this project is to predict the results of Cancer Death Rates with the following constraint, to construct an OLS regression model. Therefore, the precise and accurate prediction of ***TARGET\_deathRate*** is very important to our case. That is why we choose the ***TARGET\_deathRate*** as our target categorical variable to model with multiple regression.

## Perform Multiple Regression

To get a better regression result, we finally choose three independent variables, ***incidenceRate, medIncome, and PctPrivateCoverage***. The incidenceRate means the mean per capita (100,000) cancer diagoses(a). The medIncome means the median income per county (b). The PctPrivateCoverage means the percent of county residents with private health coverage (b).

The regression result is as follow. The R- squared is 0.404 and the adjusted R-squared is 0.403. It means that the Death Rate can be 40.4% explained by the three independent variables. Also, the 3 independent variables are statistically significant as their p-value are far less than 0.05. We can see the coefficient of Incidence Rate is 0.239 which means when the Incidence Rate increases one unit, the Death Rate will increase 0.239 unit. It makes sense as the higher Incidence Rate, the higher the Death Rate. The coefficient of Median Income is -0.000463 which means when the Median Income increases one unit, the Death Rate will decrease 0.000463 unit. The reason may be that the higher income, the more money can be used to cure cancer. Thus, the Death Rate will decline with the higher Median Income. However, the influence of Median Income is smaller than the other two variables. The coefficient of Private Coverage is -0.759 which means when the Private Coverage increases one unit, the Death Rate will decrease 0.759 unit. This may be because the more Private Coverage the people buy, the more concerned they will be about their health. Thus, the Death Rate will decline with the higher Private Coverage.



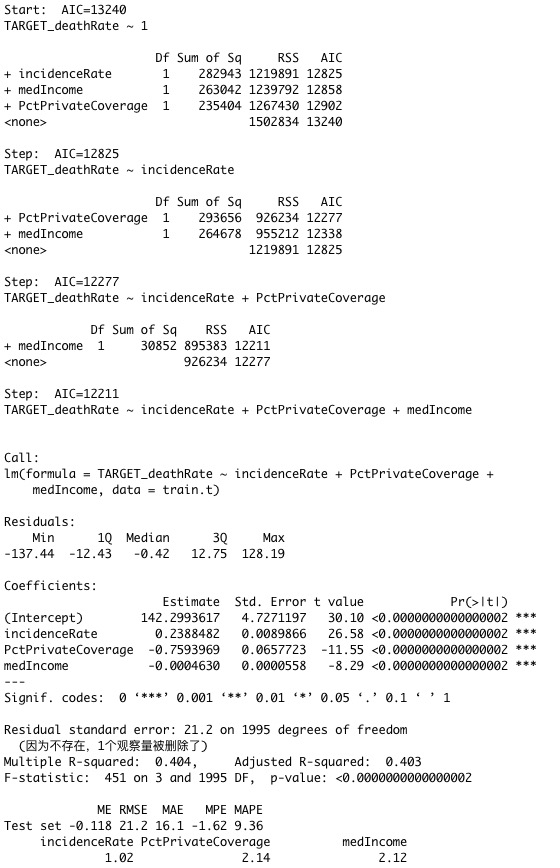
### Forward Selection

Forward selection begins with an empty model and adds in [variables](https://www.statisticshowto.com/probability-and-statistics/types-of-variables/)one by one. In each forward step, we add the one variable that gives the single best improvement to our model.

1. Begins with a model that contains no variables (called the Null Model).

2. Then starts adding the most significant variables one after the other.

3. Until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model.



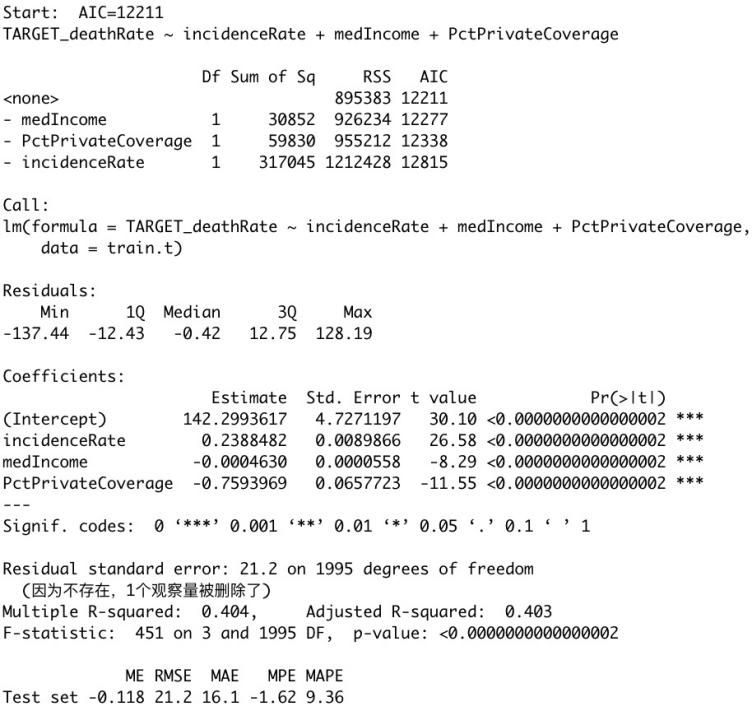
### Backward Selection

Backward selection starts with a model that includes every possible variable, and then eliminates the [extraneous variables](https://www.statisticshowto.com/extraneous-variable/) one by one.

## 1. Begins with a model that contains all variables under consideration (called the Full Model).

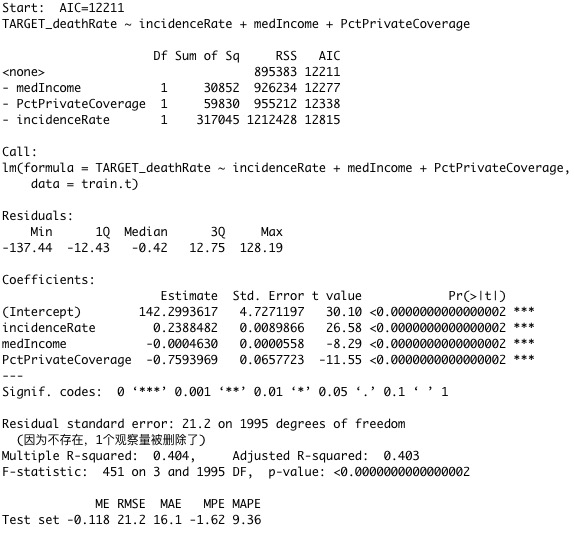
## 2. Then starts removing the least significant variables one after the other.

## 3. Until a pre-specified stopping rule is reached or until no variable is left in the model.



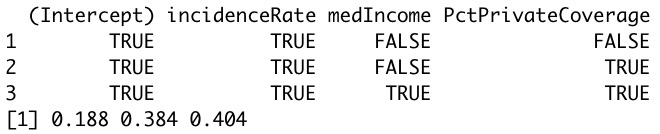
### Forward-and-Backward Selection

## Forward-and-Backward Selection, also called stepwise regression, is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Forward-and-Backward Selection is a combination of Forward selection and Backward elimination, testing at each step for variables to be included or excluded.



### Best Subset (Exhaustive) Search

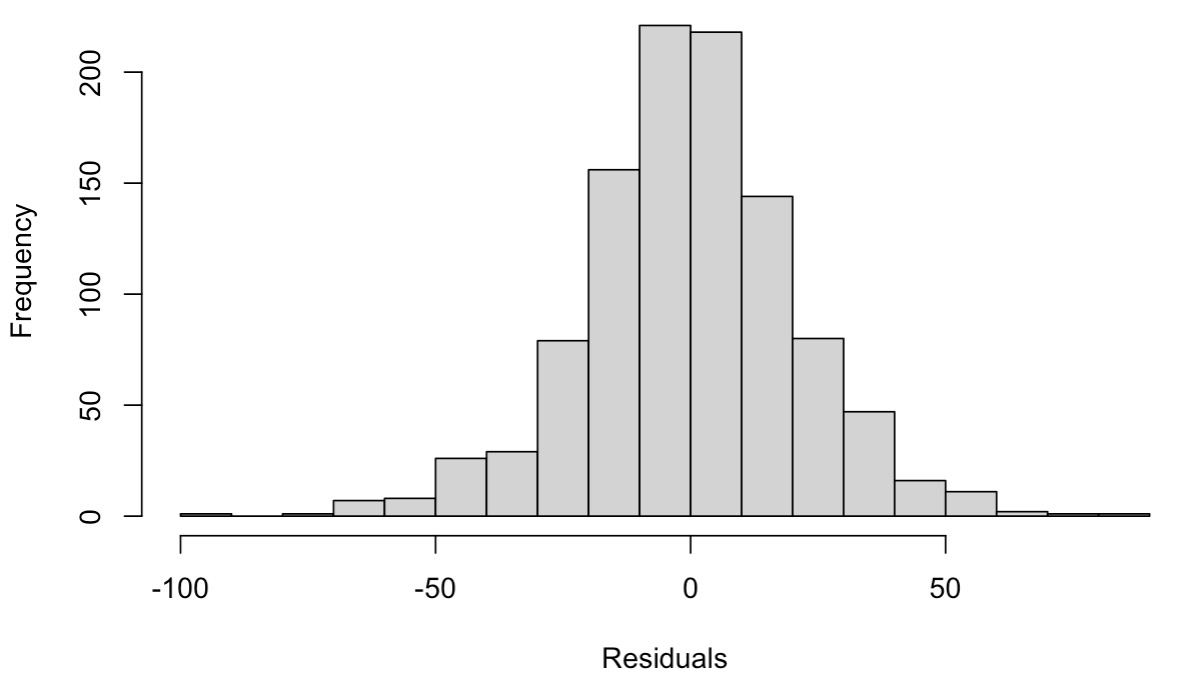
Best subset (exhaustive) search suggests generating each and every element of the problem domain, selecting those of them that satisfy all the constraints, and then finding a desired element (e.g., the one that optimizes some objective function).



## Diagnostic Residual Plots

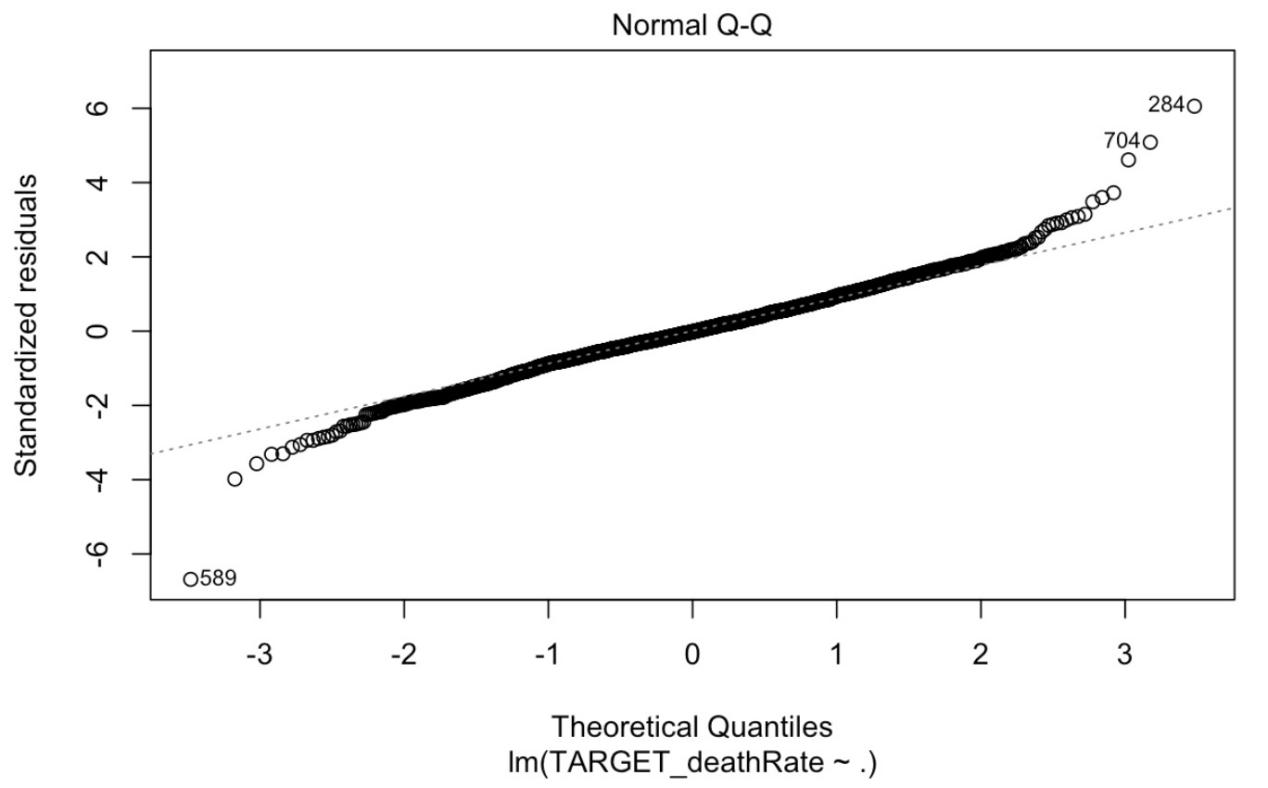
### Histogram of Residuals.

### The residuals look Gaussian, a bell-shaped curve.



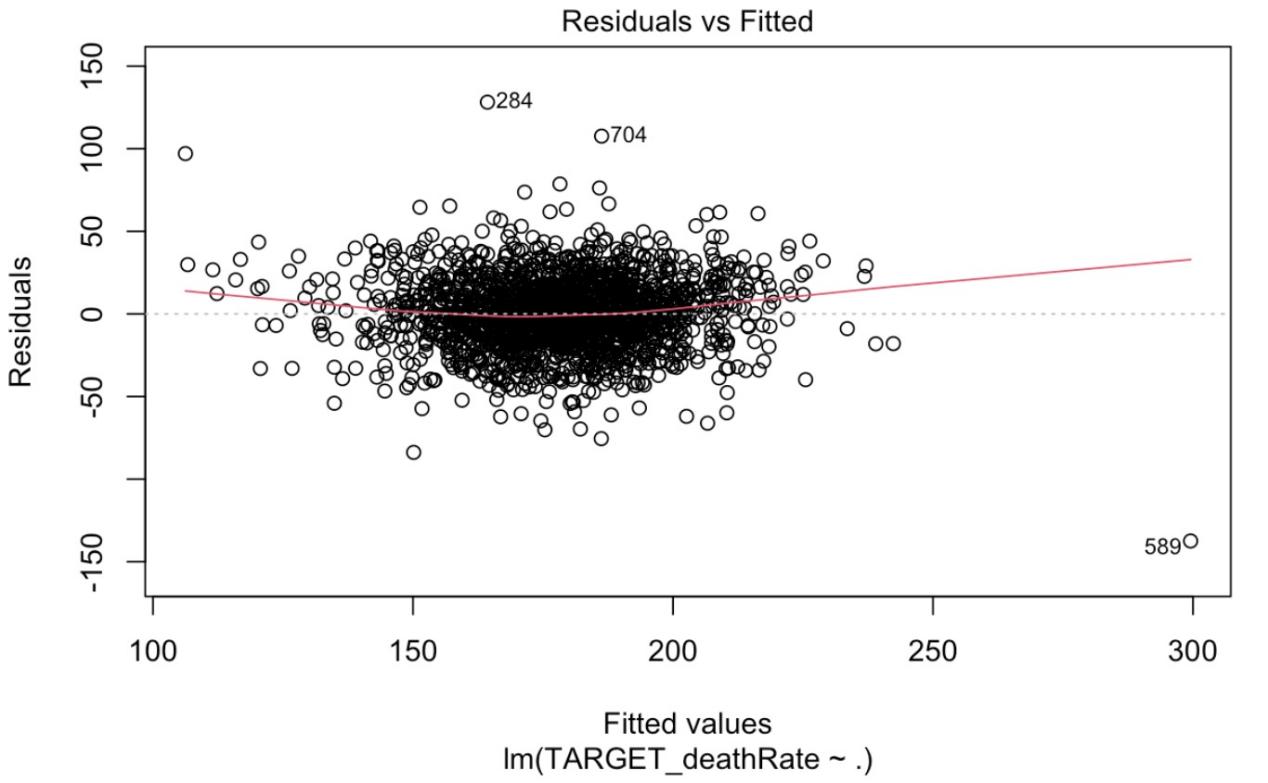
### Normal Probability Plot of Residual.

### The residuals closely track the diagonal line.



### Residuals vs. Fitted Values.

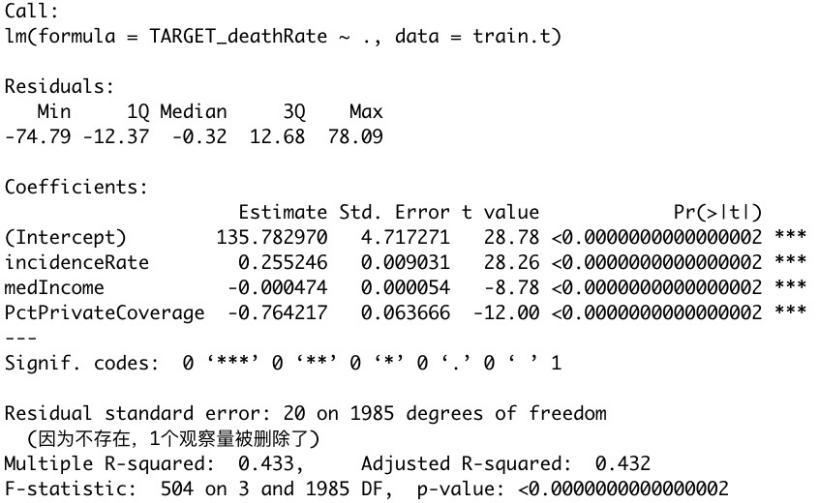
### The residuals look randomly distributed.

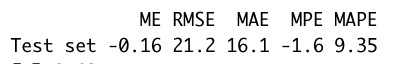


## Best Model

### Adjusted R-Squared & RMSE

### The adjusted R-Squared value of my best model is 0.432 and the RMSE is 21.2.



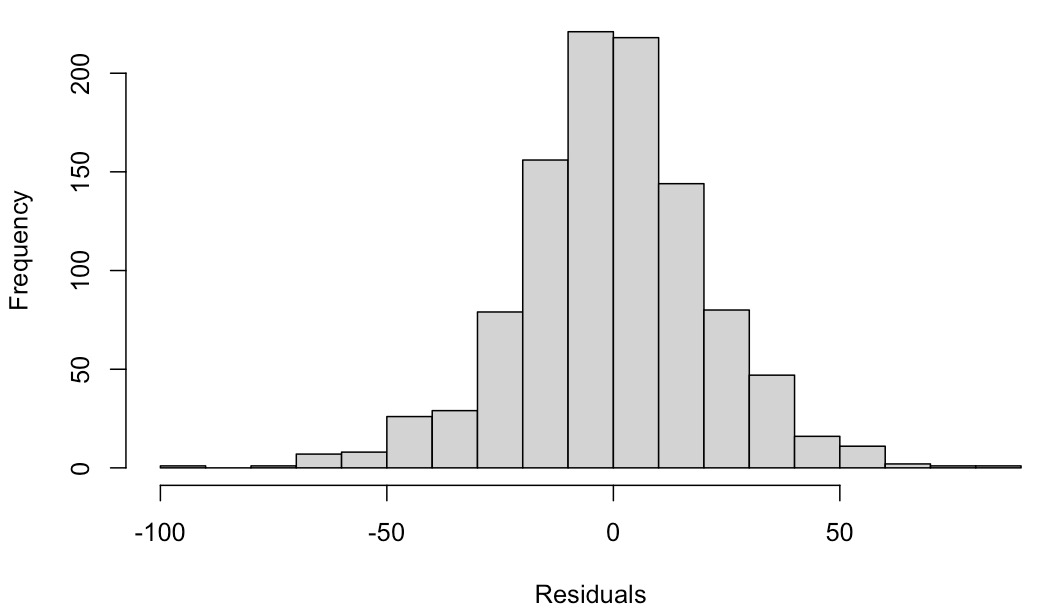


### Diagnostic Residual Plots

### I removed 10 outliers from my observations which is less than 5%. Therefore, I think it is acceptable.

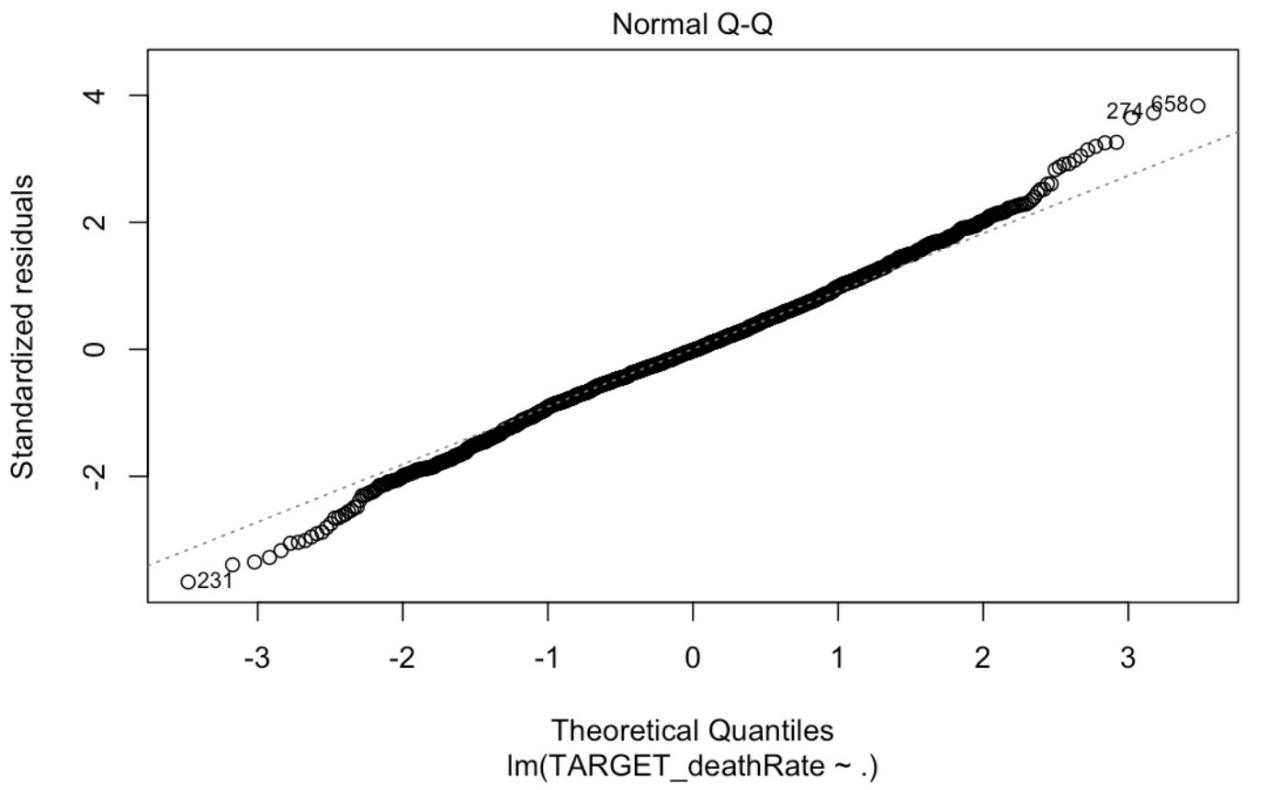
### Histogram of Residuals.

### The residuals look Gaussian, a bell-shaped curve.



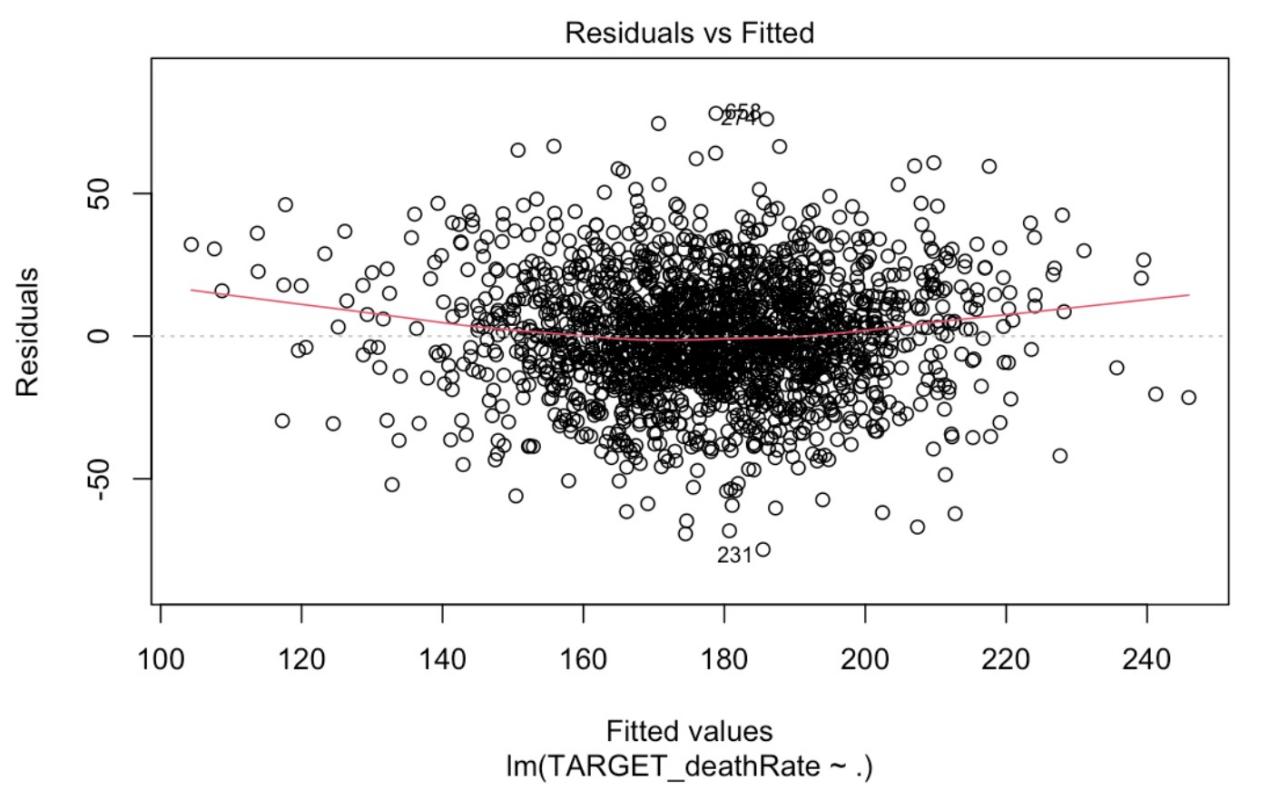
### Normal Probability Plot of Residual.

### The residuals closely track the diagonal line.



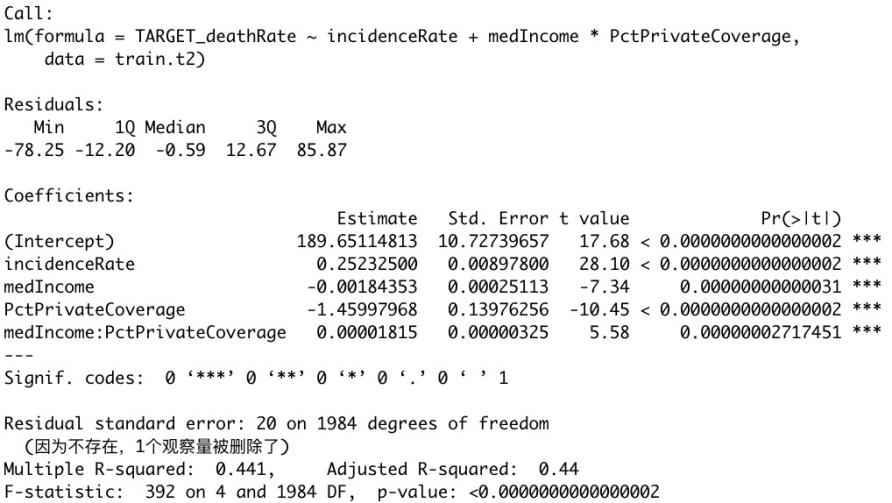
### Residuals vs. Fitted Values.

### The residuals look randomly distributed.



## Interaction Term Model

To get a better regression result, we finally choose these independent variables, ***incidenceRate, medIncome, PctPrivateCoverage, and interaction term (medIncome \* PctPrivateCoverage)***. The incidenceRate means the mean per capita (100,000) cancer diagoses(a). The medIncome means the median income per county (b). The PctPrivateCoverage means the percent of county residents with private health coverage (b). Our final adjusted R-Squared after trying to include the interaction term is 0.441 and the RMSE is 21.3. The model with interaction term performs little better than the previous model which R-Squared is 0.433 and RMSE is 21.2. The improvement of including the interaction term is not as good as imaging. The reason may be the database itself as it has a weak regression relationship with the other elements. Classification may be more suitable for this data.





The R- squared is 0.441 and the adjusted R-squared is 0.44. It means that the Death Rate can be 44.1% explained by the three independent variables. Also, the 4 independent variables are statistically significant as their p-value are far less than 0.05. We can see the coefficient of Incidence Rate is 0.252 which means when the Incidence Rate increases one unit, the Death Rate will increase 0.252 unit. It makes sense as the higher Incidence Rate, the higher the Death Rate. The coefficient of Incidence Rate is higher in the new model. The coefficient of Median Income is -0.00184 which means when the Median Income increases one unit, the Death Rate will decrease 0.00184 unit. The reason may be that the higher income, the more money can be used to cure cancer. Thus, the Death Rate will decline with the higher Median Income. The coefficient of Median Income is lower in the new model. The coefficient of Private Coverage is -1.460 which means when the Private Coverage increases one unit, the Death Rate will decrease -1.460 unit. This may be because the more Private Coverage the people buy, the more concerned they will be about their health. Thus, the Death Rate will decline with the higher Private Coverage. The coefficient of Private Coverage is lower in the new model. The coefficient of Median Income \* Private Coverage is 0.00001815 which means when the Median Income \* Private Coverage increases one unit, the Death Rate will increase 0.00001815 unit.

## Interpretation of Final Model

Cancer affects everyone, and the occurrence of cancer will cause a huge burden on patients and families. Cancer is also one of the leading causes of death in the world. However, 30%-50% of cancers are caused by long-term unhealthy lifestyle habits or poor public health measures. In this study, we screened data on many external factors to support our conclusions. We finally choose three independent variables for our regression, ***incidenceRate, medIncome, PctPrivateCoverage***. The incidenceRate means the mean per capita (100,000) cancer diagoses(a). The medIncome means the median income per county (b). The PctPrivateCoverage means the percent of county residents with private health coverage (b).

The World Health Organization (WHO) mentions several findings of cancer:

1. About 70% of cancer deaths occur in low-income or middle-income countries;
2. Less than 30% of low-income countries report public access to treatment services, compared with more than 90% in high-income countries;
3. Cancer-causing infections accounted for up to 25% of newly diagnosed cancer cases in low-income and middle-income countries in 2012;
4. In low-income and middle-income countries, only one in five have the necessary data to drive cancer policy.

The results of WHO are proved in our findings. We can see the coefficient of Median Income is positive which means the Death Rate will decline with the higher Median Income. This may reason as the higher income, the more money can be used to cure cancer. Also, The Death Rate will decline with the higher Private Coverage. The more Private Coverage the people buy, the richer they will be and the more concerned they will be about their health. In places with higher per capita income, people have more money to care for and protect their health. They have regular checkups and healthy assurance before they get sick, and they have timely, professional, and high-tech treatments after they get sick. Therefore, their cancer mortality rate will be much lower than that of ordinary people.

Cancer is increasingly a microcosm of inequality. On the one hand, with the increasing popularity of cancer screening for the purpose of early detection of lesions, coupled with the improvement of treatment methods, cancer patients are living longer than ever before. But these advances have not benefited everyone equally. As the gap between the rich and the poor widens, the gap between the rich and the poor in cancer mortality will continue to increase. In 2017, the Journal of the Environment and Public Health published an article on the impact of different socioeconomic and racial conditions on cancer mortality. The researchers found that relatively affluent, well-educated, and high-income areas of the United States had lower cancer mortality rates than relatively poorer, less-educated areas. Based on more effective prevention strategies, earlier detection of cancer and more high-quality medical resources, most cancer-related mortality rates will be effectively reduced. This also illustrates the important role of government departments in cancer. By increasing the health care coverage of the poor and making them more accessible to medical services, including cancer diagnosis and treatment, the gap between rich and poor in cancer mortality may slowly narrow.