Introduction

The main goal of our project was to evaluate the predictive impact of 6 or more chosen variables on cancer death rate. To achieve this, I chose variables that indicate a person's socioeconomic status, demographic information, and healthcare coverage. Varying insurance policies was a subject I wanted to dive into, given the large discourse around Healthcare in the United States.

Variable Selection:

`povertyPercent`
`PctPublicCoverage`
`PctUnemployed16_Over`
`PctPrivateCoverage`
`incidenceRate`
`medIncome`
`MedianAge`

General Overview

1. Variable Selection
2. Exploratory Data Analysis
3. Create Model
4. Model Selection
5. Diagnostic Tests

Exploratory Data Analysis

This step in the process is used to identify any issues within the data that may lead us to produce inaccurate or skewed results. The main theme we wanted to look out for was: Multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated to one another. This leads to inaccurate estimates and poor performing models.

1. **Correlation Matrix**

Using the Correlation Matrix method, I was able to identify several predictor variables that were strongly correlated to one another.

Poverty Percent <–> Private Coverage Percentage
Poverty Percent <–> Median Income
Public Coverage Percentage <–> Median Income

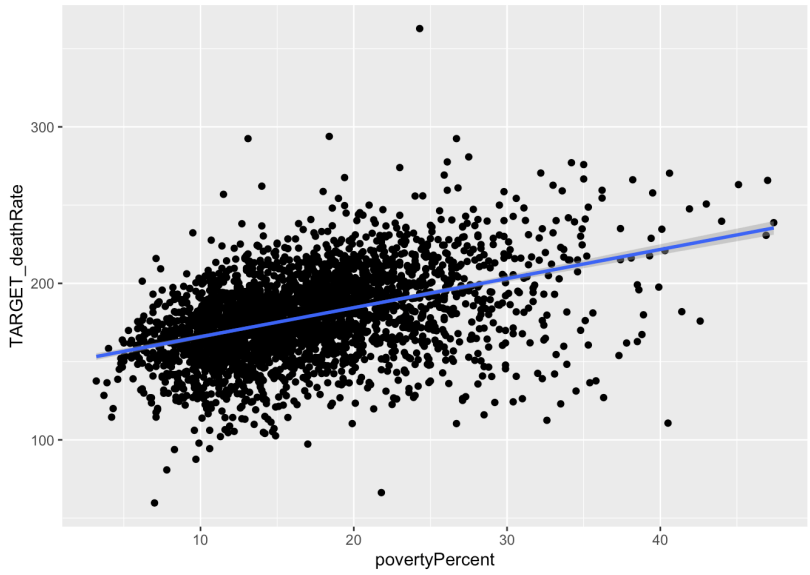
Public Coverage Percentage <–> Private Coverage Percentage

Highly correlated variables lead to multicollinearity in the model, which we want to avoid.

2. Scatter Plots

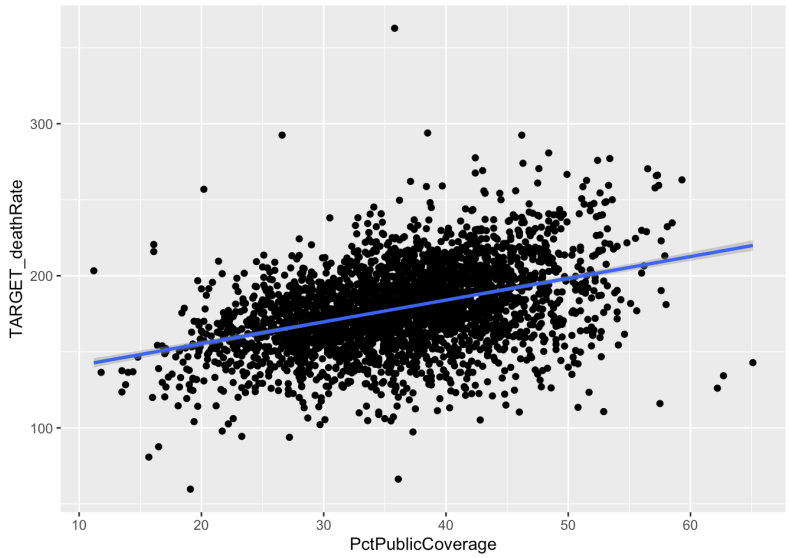
The next step in our EDA process is to examine the relationship between each predictor variable and the target variable.

Poverty Percent



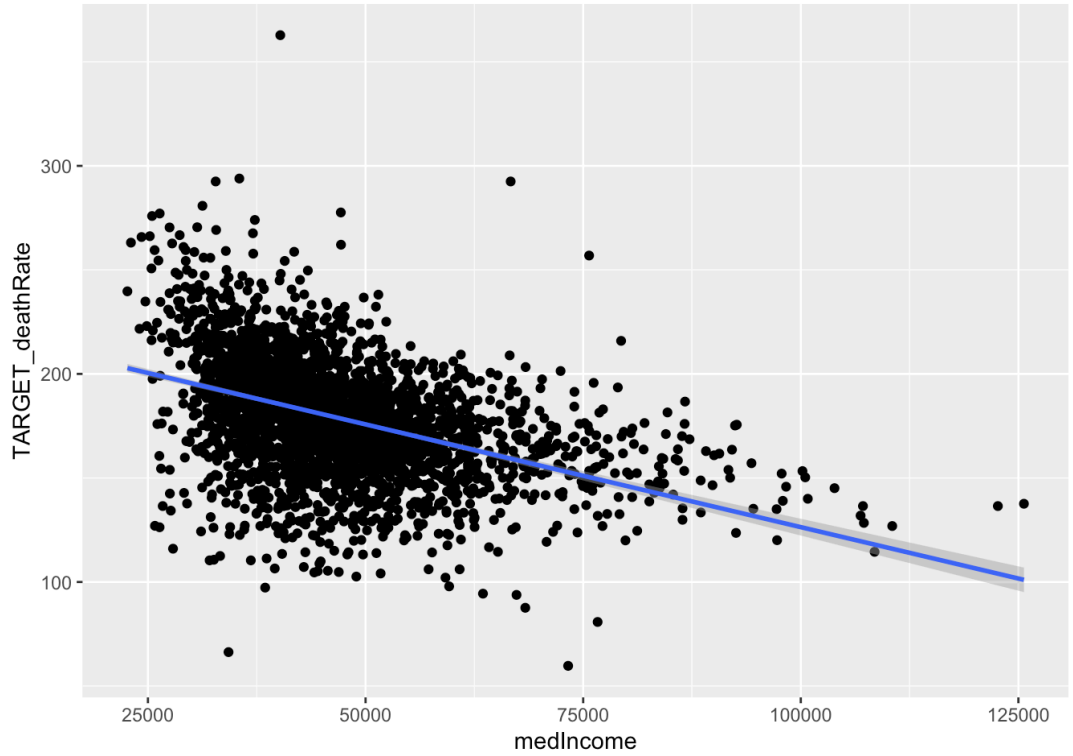
Poverty Percentage has a positive correlation with TARGET_deathRate. The higher the poverty rate, the higher the deathRate.

Public Coverage Percentage



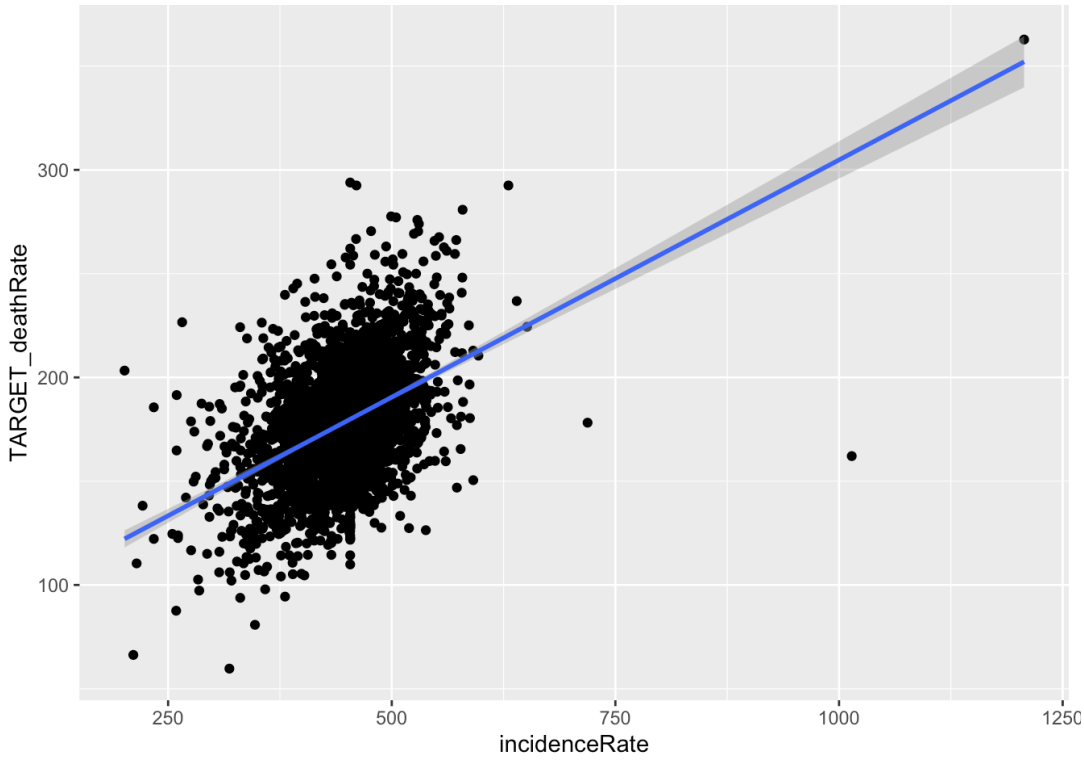
Public Coverage Percentage has a positive correlation with TARGET_deathRate. The higher the Public Coverage Percentage rate, the higher the TARGET_deathRate. This graph does have less variance than the Poverty Percentage graph.

Median Income



Median Income has a negative correlation with TARGET_deathRate. The higher the Median Income, the lower the deathRate.

Incidence Rate

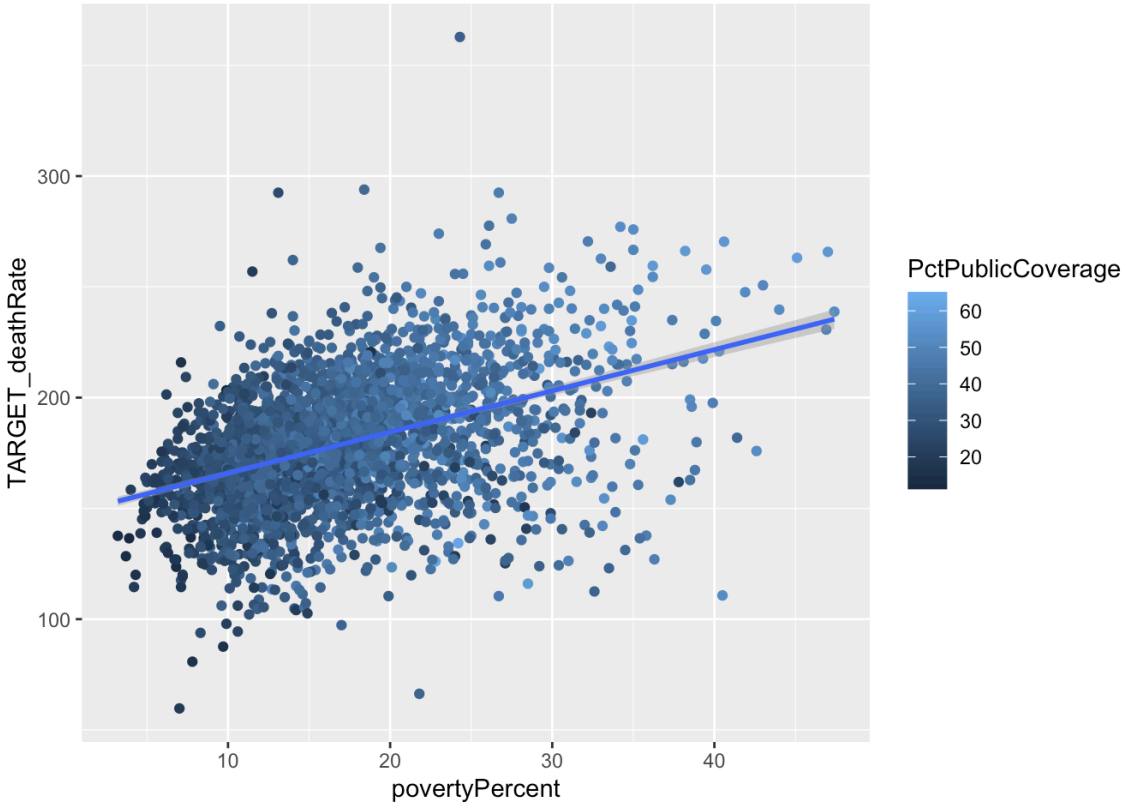


Incidence Rate has a positive correlation with TARGET_deathRate. The higher the Incidence Rate, the higher the deathRate. This graph did have the most variance.

3. Multivariate Scatter Plot

Similar to scatter plots, but with 3 or more variables plotted, as opposed to just 2. Here, I chose:

Public Coverage Percentage and Poverty Percent



Poverty Percentage has a positive correlation with TARGET_deathRate. When Public Coverage Percentage increases, both the deathRate and Poverty Percentage increases.

Create Model

```
model <- lm(formula1, data=cancer_df)
summary(model)
```

```
Call:
lm(formula = formula1, data = cancer_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-120.161  -13.330   -0.116   13.425  173.061 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.433e+02  9.965e+00  14.382  < 2e-16 ***
povertyPercent      6.348e-01  1.401e-01   4.532 6.06e-06 ***
PctPublicCoverage   1.047e+00  9.729e-02  10.767  < 2e-16 ***
PctEmpPrivCoverage  1.344e+00  9.614e-02  13.975  < 2e-16 ***
PctPrivateCoverage -6.466e-01  9.417e-02  -6.866 7.96e-12 ***
studyPerCap        -5.462e-04  8.215e-04  -0.665    0.506    
medIncome          -5.789e-04  7.083e-05  -8.173 4.38e-16 ***
MedianAge           5.259e-03  9.585e-03   0.549    0.583    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.83 on 3039 degrees of freedom
Multiple R-squared:  0.2641,    Adjusted R-squared:  0.2624 
F-statistic: 155.8 on 7 and 3039 DF,  p-value: < 2.2e-16
```

Using our selected variables, we create a linear regression model. Printing out the summary of the model, we can see which variables are deemed statistically significant as a predictor for the death rate.

- PctPublicCoverage and PctEmpPrivCoverage have positive slope estimates, which means for every unit increase in PctPublicCoverage or PctEmpPrivCoverage, the death rate increases. Prior to the model, I would have expected insurance coverage of any kind to have a negative correlation.
- povertyPercent, PctPublicCoverage, PctEmpPrivCoverage, PctPrivateCoverage are statistically significant. I would have predicted that MedianAge would be statistically significant as, from personal experience, the risk of cancer increases with age.
- **R-squared:** 0.2641. This means that the predictors chosen only account for 26.41% of the variance in TARGET_deathRate.

Model Selection

Next, we perform model selection between fastbw and stepAIC. Both models aim to eliminate variables with low predictive power in regards to the death rate. However, they use different algorithms and thresholds to make these selections. We must choose which model we are more comfortable moving forward with.

1. fastbw()

```
ols.mod1 <- ols(formula1, data=cancer_df)
fastbw(ols.mod1, rule="p", sls=0.05)
```

```
 Deleted     Chi-Sq d.f. P      Residual d.f. P      AIC   R2
 MedianAge   0.30   1    0.5832 0.30     1    0.5832 -1.70 0.264
 studyPerCap 0.46   1    0.4975 0.76     2    0.6834 -3.24 0.264

Approximate Estimates after Deleting Factors

                       Coef      S.E. Wald Z         P
Intercept           1.438e+02 9.9515368 14.448 0.000e+00
povertyPercent      6.300e-01 0.1398885  4.504 6.676e-06
PctPublicCoverage   1.048e+00 0.0971216 10.793 0.000e+00
PctEmpPrivCoverage  1.338e+00 0.0958992 13.949 0.000e+00
PctPrivateCoverage -6.475e-01 0.0940836 -6.883 5.879e-12
medIncome          -5.777e-04 0.0000708 -8.160 3.331e-16

Factors in Final Model

[1] povertyPercent     PctPublicCoverage  PctEmpPrivCoverage PctPrivateCoverage
[5] medIncome
```

MedianAge and studyPerCap both did not have a significant p value in the previous model. Actually they both have the highest p values, and the fastbw method reflects that.

2. stepAIC()

```
aic_result <- stepAIC(model)
```

```
Step:  AIC=19329.63
TARGET_deathRate ~ povertyPercent + PctPublicCoverage + PctEmpPrivCoverage +
    PctPrivateCoverage + medIncome

                     Df Sum of Sq     RSS   AIC
<none>                             1726833 19330
- povertyPercent      1     11523 1738356 19348
- PctPrivateCoverage  1     26910 1753742 19375
- medIncome           1     37826 1764658 19394
- PctPublicCoverage   1     66181 1793014 19442
- PctEmpPrivCoverage  1    110542 1837374 19517
```

MedianAge and studyPerCap both did not have a significant p value in the previous model and the stepAIC method reflects that. Those were the only two predictors removed from the model.

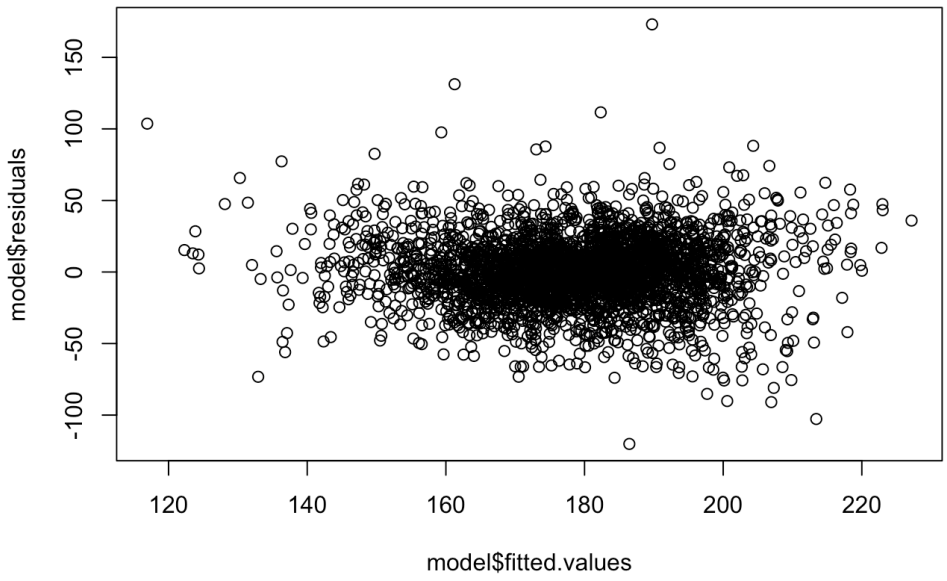
I chose to continue with the stepAIC model.

```
summary(aic_result)
```

```
Call:
lm(formula = TARGET_deathRate ~ povertyPercent + PctPublicCoverage +
    PctEmpPrivCoverage + PctPrivateCoverage + medIncome, data = cancer_df)

Residuals:
     Min       1Q   Median       3Q      Max
-120.183  -13.289   -0.142   13.401  173.157

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.438e+02  9.950e+00  14.451  < 2e-16 ***
povertyPercent      6.300e-01  1.399e-01   4.505 6.90e-06 ***
PctPublicCoverage   1.048e+00  9.710e-02  10.796  < 2e-16 ***
PctEmpPrivCoverage  1.338e+00  9.588e-02  13.952  < 2e-16 ***
PctPrivateCoverage -6.475e-01  9.406e-02  -6.884 7.04e-12 ***
medIncome          -5.777e-04  7.079e-05  -8.162 4.79e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.83 on 3041 degrees of freedom
Multiple R-squared:  0.2639,    Adjusted R-squared:  0.2627
F-statistic:   218 on 5 and 3041 DF,  p-value: < 2.2e-16
```

Diagnostic Tests

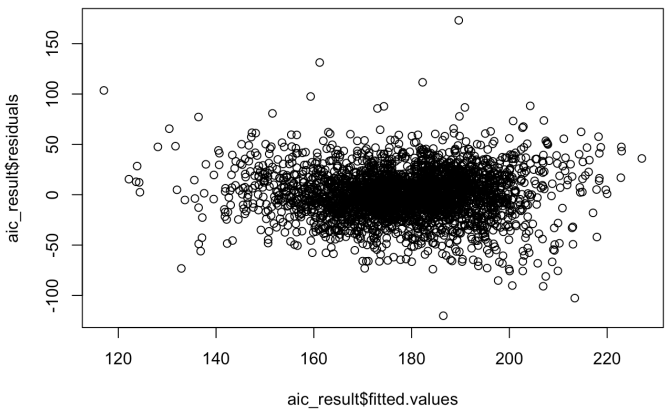
```
plot(model$fitted.values, model$residuals)
```



The residuals are clustered at the center, but increase in spread at the extremes. This may indicate slight heteroscedasticity, which is a violation of the constant variance assumption.

1. **Fitted Values vs Residuals Plot**

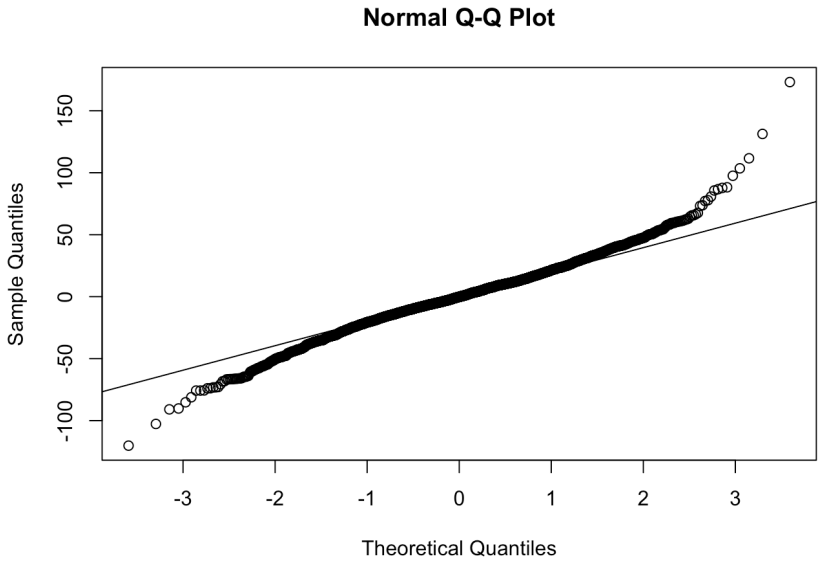
```
plot(aic_result$fitted.values, aic_result$residuals)
```



The residuals are clustered at the center, but increase in spread at the extremes. This may indicate slight heteroscedasticity, which is a violation of the constant variance assumption.

2. Q-Q Plot

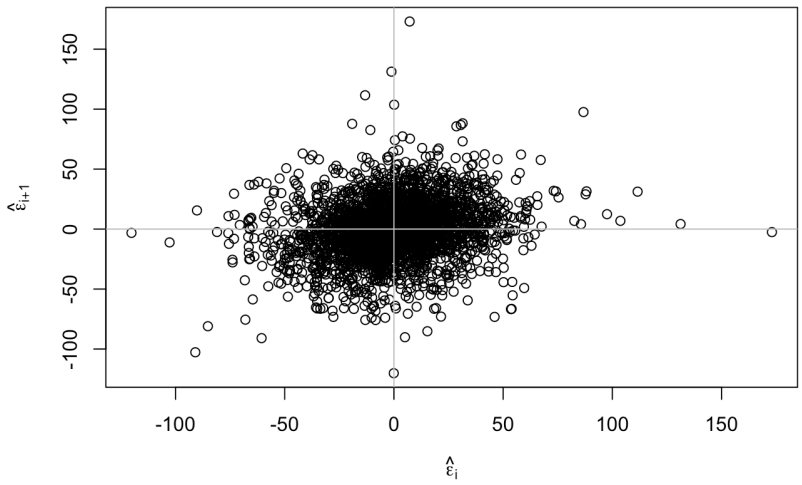
```
qqnorm(aic_result$residuals)
qqline(aic_result$residuals)
```



There are quite a number of points that deviate from the line. There is evidence for non normal residuals

3. Lagged Residuals Plot

```
n <- length(residuals(model))
plot(tail(residuals(model),n-1) ~ head(residuals(model),n-1), xlab=expression(hat(epsilo
abline(h=0,v=0,col=grey(0.75))
```



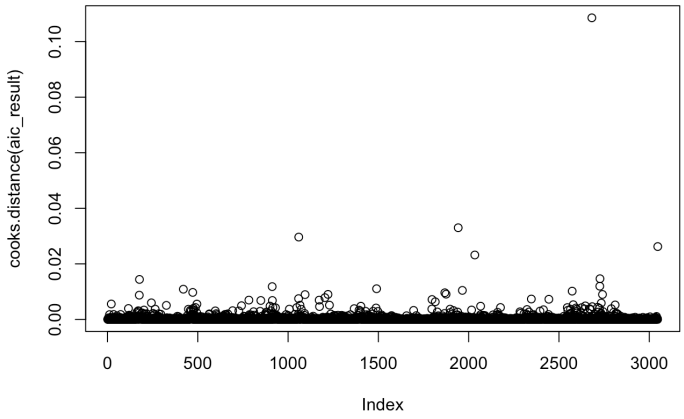
The lagged residual plot does not show a linear relationship between consecutive residuals. One residual does not strongly predict the next.

Evaluating Outliers

1. **Cook's Distance**

Identifies points that disproportionately affect the model's coefficients.

```
plot(cooks.distance(aic_result))
```

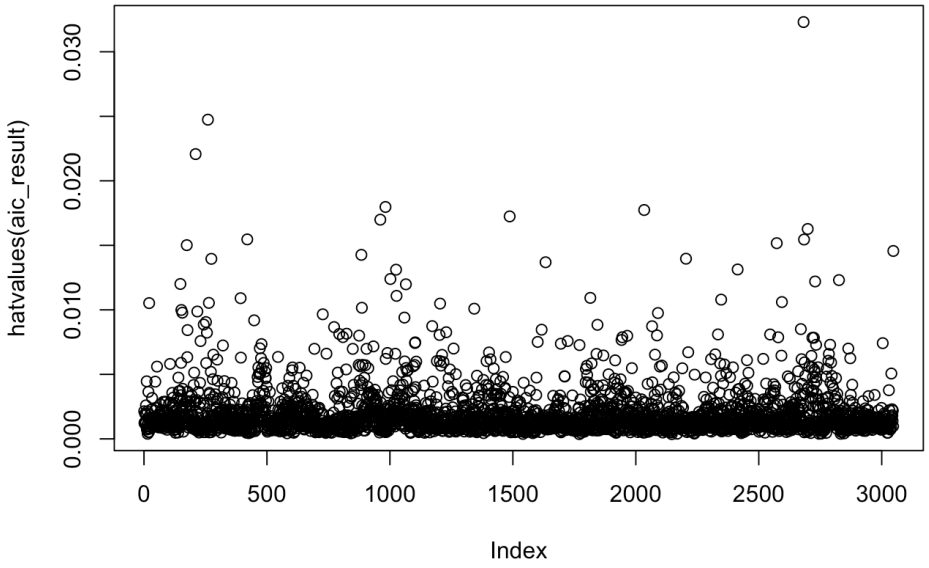


From observing the Cook's distance plot, we can see one clear outlier. Most are very close to 0, very low Cook's distance numbers.

2. **Hat Values**

Shows how far an individual predictor value is from the mean of all predictor values.

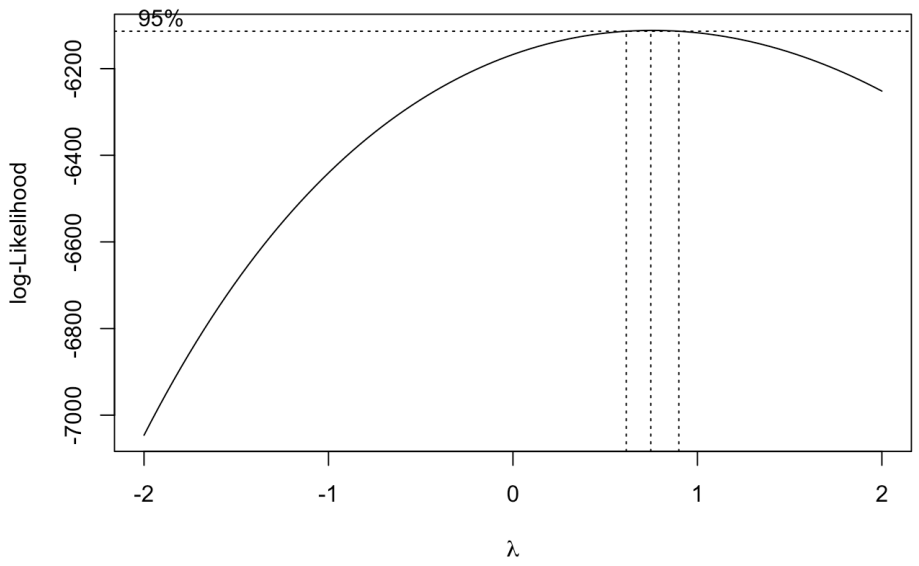
```
plot(hatvalues(aic_result))
```



BoxCox

From the information gathered in the Diagnostic Tests, we conclude that a transformation is needed. To correct this, I will use a Box-Cox transformation

```
bc <- boxcox(model, plotit=T)
```




```
lambda <- bc$x[which.max(bc$y)]
lambda
```

```
[1] 0.7474747
```

This value of lambda indicates that we need a power transformation.

```r
formula2 <- as.formula(TARGET_deathRate^lambda ~ povertyPercent + PctPublicCoverage + Pc
modelBox <- lm(formula2, data=cancer_df)
summary(modelBox)
```

```
Call:
lm(formula = formula2, data = cancer_df)

Residuals:
     Min       1Q   Median       3Q      Max
-26.6464  -2.6620   0.0447   2.7795  31.5744

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.101e+01  2.014e+00  20.366  < 2e-16 ***
povertyPercent      1.200e-01  2.830e-02   4.241 2.29e-05 ***
PctPublicCoverage   2.147e-01  1.966e-02  10.920  < 2e-16 ***
PctEmpPrivCoverage  2.773e-01  1.943e-02  14.277  < 2e-16 ***
PctPrivateCoverage -1.315e-01  1.903e-02  -6.909 5.90e-12 ***
studyPerCap        -1.037e-04  1.660e-04  -0.625    0.532
medIncome          -1.209e-04  1.431e-05  -8.446  < 2e-16 ***
MedianAge           1.030e-03  1.937e-03   0.532    0.595
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.816 on 3039 degrees of freedom
Multiple R-squared:  0.263, Adjusted R-squared:  0.2613
F-statistic: 154.9 on 7 and 3039 DF,  p-value: < 2.2e-16
```

Final Results

P-values

```r
summary_model <- summary(modelBox)
coef <- summary_model$coefficients
coef_df <- as.data.frame(coef)
coef_df <- coef_df[, c("Estimate", "Pr(>|t|)")]
colnames(coef_df) <- c("Parameter Estimate", "p-value")
print(coef_df)
```

	Parameter Estimate	p-value
(Intercept)	41.0079352787	1.565360e-86
povertyPercent	0.1200135862	2.294789e-05
PctPublicCoverage	0.2146638982	2.952828e-27
PctEmpPrivCoverage	0.2773399677	8.332165e-45
PctPrivateCoverage	-0.1314676344	5.901611e-12
studyPerCap	-0.0001036752	5.322672e-01
medIncome	-0.0001208764	4.608306e-17
MedianAge	0.0010304328	5.947387e-01

As shown above, only 2 predictors had a p-value greater than our threshold value of 0.05, which re MedianAge and studyPerCap.

R-squared

```
summary_model$r.squared
```

```
[1] 0.2629508
```

This means that 26.3% of the variance in the model can be explained by the variables chosen.

Conclusion

We began this project with a dataset of cancer data, which included a death rate (response variable) and numerous other predictor variables. From that, I chose 7 variables I was interested in analyzing, where I focused on variables that were indicators of socioeconomic status, demographic information, and healthcare coverage. From those variables, there were 2 that were deemed statistically insignificant: studyPerCap and MedianAge.

The remaining 5 were deemed significant, though 2 stood out as being highly statistically significant: PctPublicCoverage and PctEmpPrivCoverage. What is really interesting about this is that they have positive slope estimates, which means that higher rates of both variables lead to higher death rates.

These results at a glance may seem counterintuitive, or perhaps these results are correct and having these forms of health coverage does increase the death rate. Regardless, the study has generated fascinating insights and warrants even further investigation.