Predictive Modeling of Cancer Death Ratess

AUTHOR
Cristian T

PUBLISHED
August 6, 2024

Predictive Modeling of Cancer Death Rates Using Demographic and Socioeconomic Factors

Cancer remains a significant public health challenge. Various demographic and socioeconomic factors, including median income, population estimates, poverty percentage, average household size, percentage of married individuals, employment rates, education levels, and health insurance coverage, influence mortality rates. This project utilizes a comprehensive dataset aggregated from multiple sources, such as the American Community Survey, clinicaltrials.gov, and cancer.gov, covering a variety of US counties. The dataset includes a wealth of health-related information and socioeconomic data, providing a robust foundation for our analysis. Understanding these drivers is crucial for developing effective interventions and policies to reduce cancer's impact on communities. This project aims to develop a predictive model that uses these variables to forecast cancer death rates accurately. We will employ data-driven approaches to identify key predictors and explore the relationships between demographic characteristics, socioeconomic status, and cancer outcomes.

We begin by loading the project dataset containing 30 socioeconomic and health-related variables. Next, we create a subset of the data containing only the variables relevant to our analysis. These variables include median income, population estimates, poverty percentages, and other socioeconomic indicators. To gain insights into the structure and summary statistics of the filtered dataset, we utilize the skim() function from the skimr package. Among the notable findings, the mean cancer death rate is approximately 179 deaths per 100,000 people, with a considerable standard deviation of 27.8, indicating substantial variability in mortality rates. Median income levels exhibit a similar pattern, with a mean of $47,063 and a standard deviation of $12,040, suggesting diverse economic conditions among counties. Additionally, the estimated population in 2015 is broad, from a few hundred to over ten million people, with a mean of approximately 102,637 and a large standard deviation of 329,059. Socioeconomic indicators such as poverty percentage, average household size, and education levels vary significantly across counties, underscoring the complex interplay between economic status and health outcomes.

Data source https://www.kaggle.com/code/meenaaa/cancer-death-rate-xgboost

```
# Load necessary libraries for data analysis and visualization
suppressMessages(library(readr)) # For reading and parsing data files into R.
suppressMessages(library(skimr))  # For summary statistics and data visualization
suppressMessages(library(tidyverse))  # For data manipulation and visualization
suppressMessages(library(dplyr))  # For data manipulation
suppressMessages(library(ggplot2))  # For advanced data visualization
suppressMessages(library(stringr))  # For string manipulation functions
suppressMessages(library(MASS))  # For statistical analysis and modeling
suppressMessages(library(faraway))  # For regression diagnostics and analysis
suppressMessages(require(rms)) # For regression modeling



# Attach the 'project' dataframe to the search path
```

```
attach(project)

#limit just to the variables needed in this analysis
filt_project <- dplyr::select(project, TARGET_deathRate, medIncome, popEst2015, pover

# Print summary statistics
skim(filt_project)
```

Name	filt_project
Number of rows	3047
Number of columns	11
_____	
Column type frequency:	
numeric	11
_____	
Group variables	None

Data summary

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	
TARGET_deathRate	0	1	178.66	27.75	59.70	161.20	178.1	195.20	
medIncome	0	1	47063.28	12040.09	22640.00	38882.50	45207.0	52492.00	1
popEst2015	0	1	102637.37	329059.22	827.00	11684.00	26643.0	68671.00	101
povertyPercent	0	1	16.88	6.41	3.20	12.15	15.9	20.40	
AvgHouseholdSize	0	1	2.48	0.43	0.02	2.37	2.5	2.63	
PercentMarried	0	1	51.77	6.90	23.10	47.75	52.4	56.40	
PctNoHS18_24	0	1	18.22	8.09	0.00	12.80	17.1	22.70	
PctHS18_24	0	1	35.00	9.07	0.00	29.20	34.7	40.70	
PctBachDeg18_24	0	1	6.16	4.53	0.00	3.10	5.4	8.20	
PctEmpPrivCoverage	0	1	41.20	9.45	13.50	34.50	41.1	47.70	
PctPublicCoverage	0	1	36.25	7.84	11.20	30.90	36.3	41.55	

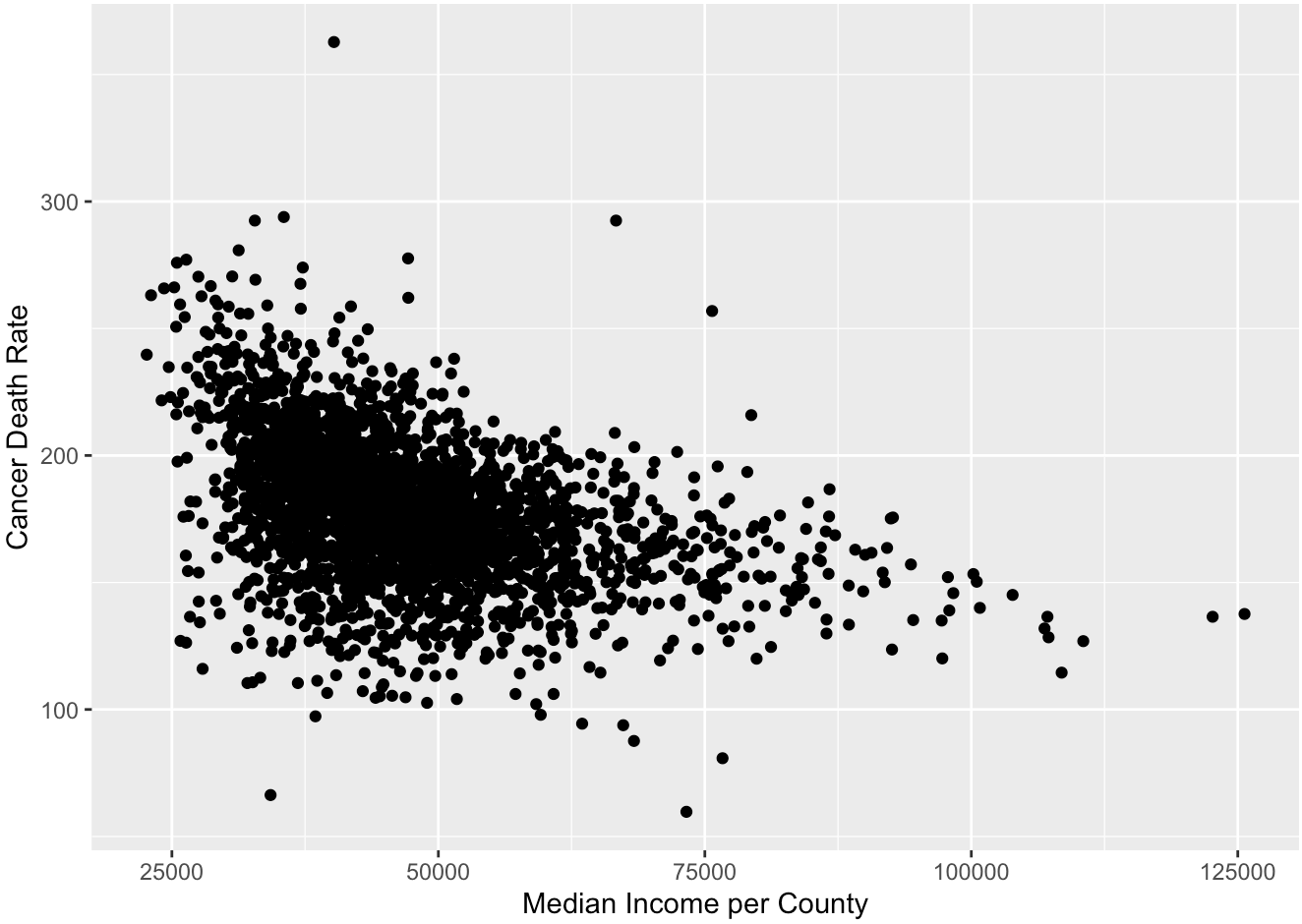
Exploratory Data Analysis (EDA)

To further explore the data, we visualized key variables using scatter plots to understand their relationships with cancer death rates.

Median Income vs. Cancer Death Rate:

```
# Scatterplot Median Income and Death Rate
ggplot(filt_project, aes(x = medIncome, y = TARGET_deathRate)) +
  geom_point() +
```

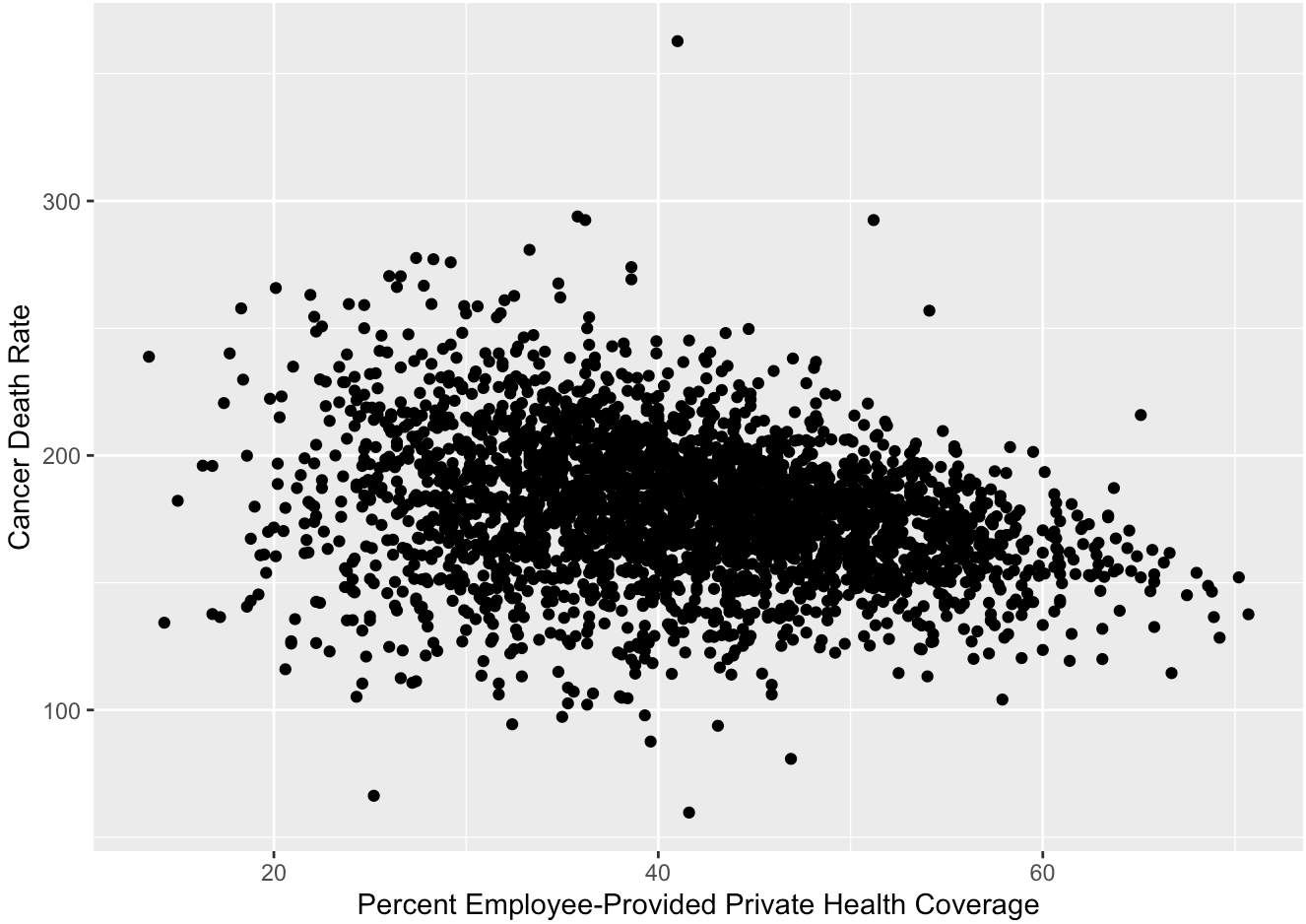
```
  labs(x = "Median Income per County",
       y = "Cancer Death Rate")
```



The plot shows a negative linear relationship, suggesting that higher median incomes, which range from $22,640 to $125,635, are associated with lower cancer death rates. Most counties cluster between $38,882 and $52,492, indicating a moderate income bracket. Notably, a few outliers suggest unique county characteristics that may influence this trend.

Employee-Provided Health Coverage vs. Cancer Death Rate

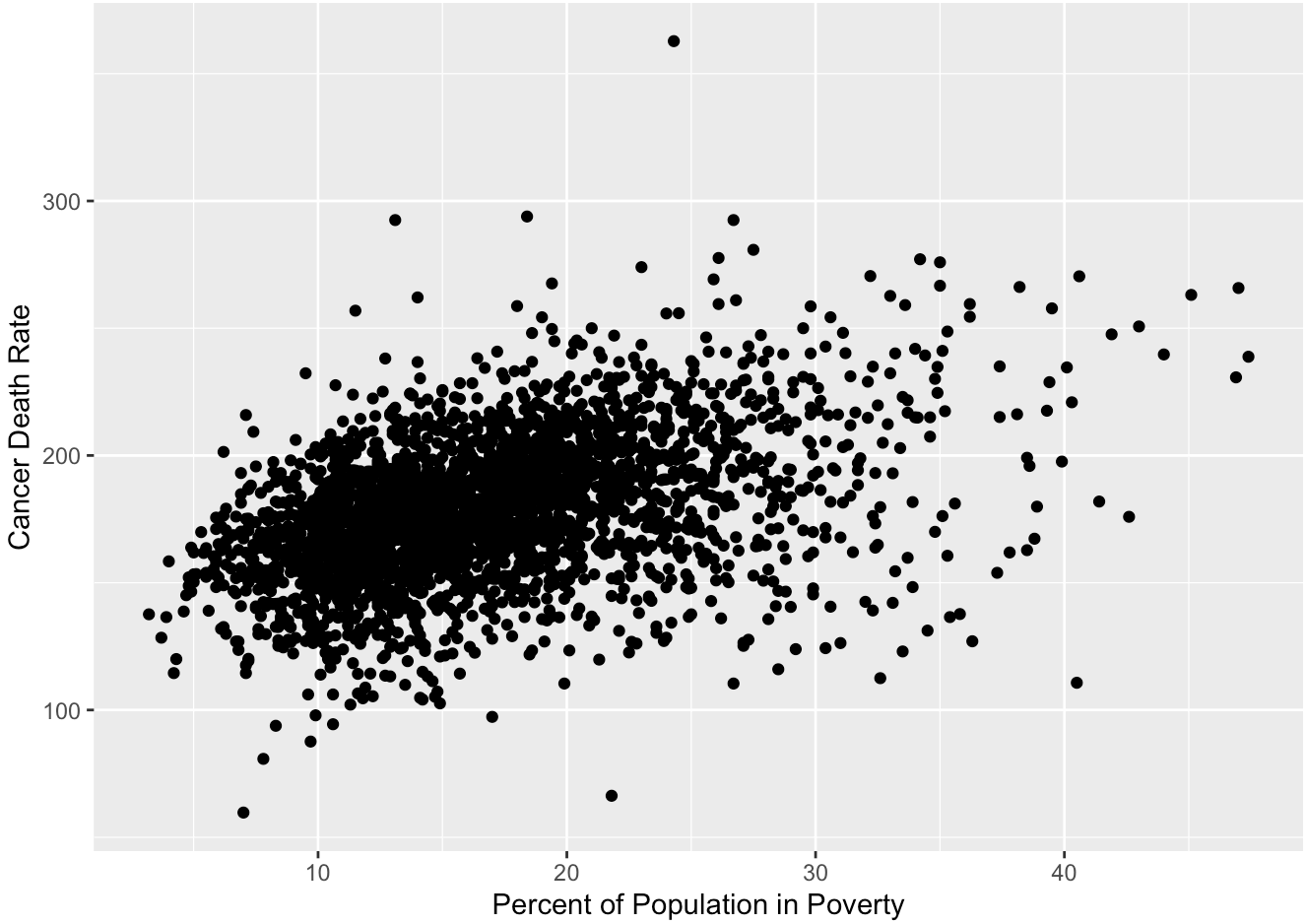
```
# Scatter plot Private HC and Death Rate
ggplot(filt_project, aes(x = PctEmpPrivCoverage, y = TARGET_deathRate)) +
  geom_point() +
  labs(x = "Percent Employee-Provided Private Health Coverage",
       y = "Cancer Death Rate")
```



This plot reveals a slight negative linear relationship, with coverage levels varying widely from 13.5% to 70.7%. The scattered data points suggest variability across counties, with some outliers that could be crucial for further investigation.

Poverty Percentage vs. Cancer Death Rate

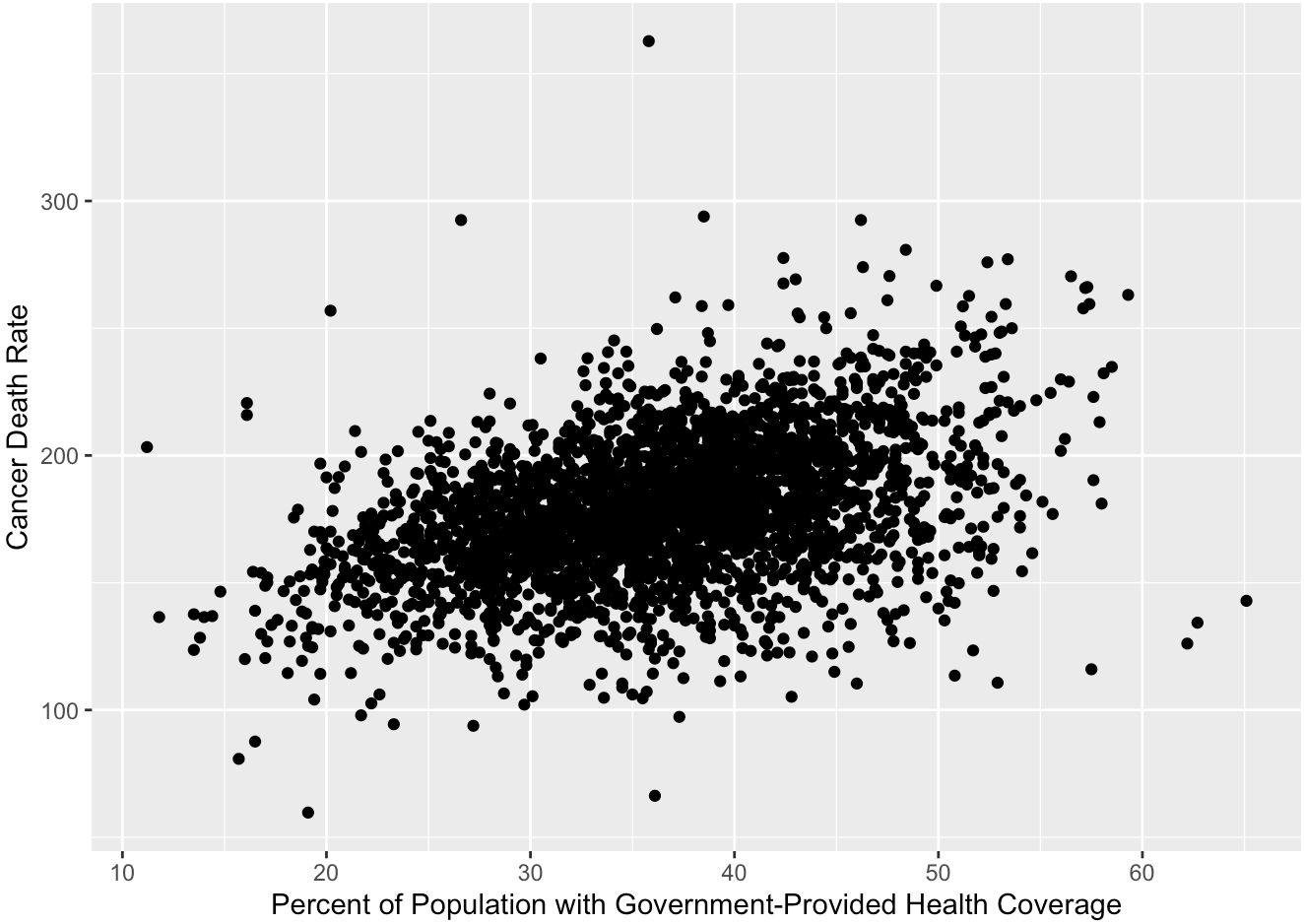
```
# Scatter plot Poverty percentage and Death Rate
ggplot(filt_project, aes(x = povertyPercent, y = TARGET_deathRate)) +
  geom_point() +
  labs(x = "Percent of Population in Poverty",
       y = "Cancer Death Rate")
```



A positive linear relationship is observed, with poverty percentages ranging from 3.20% to 47.40%. The wide spread of data points highlights significant socio-economic disparities across counties.

Government-Provided Health Coverage vs. Cancer Death Rate:

```
ggplot(project, aes(x = PctPublicCoverage, y = TARGET_deathRate)) +
  geom_point() +
  labs(x = "Percent of Population with Government-Provided Health Coverage",
       y = "Cancer Death Rate")
```



Similarly, this plot shows a positive linear relationship. Coverage ranges from 11.20% to 65.10%, with a diverse spread of data points indicating the complexity of factors influencing cancer mortality rates.

It is important to note that each relationship, while indicative of trends, requires careful interpretation to avoid conflating correlation with causation. Further analysis is needed to understand the underlying factors and the impact of outliers observed in the plots.

Fit a Linear Model

The coefficients section presents the estimated coefficients for each predictor variable in the model and their standard errors, t-values, and corresponding p-values. The standard errors reflect the variability of the coefficient estimates, indicating their precision. T-values test the significance of each coefficient, with higher values suggesting stronger evidence against the null hypothesis of no effect. P-values assess the statistical significance, with values below 0.05 typically indicating that the results are statistically significant. These coefficients represent the estimated change in the target variable for a one-unit increase in each predictor, holding all other predictors constant. For example, a negative coefficient for Median Income (medIncome) suggests that higher median income is associated with lower cancer death rates. In contrast, a positive coefficient for Poverty Percentage (povertyPercent) indicates that higher poverty percentages are associated with higher cancer death rates. The model also reveals several significant predictors that are associated with cancer death rates, including “medIncome”, “povertyPercent”, Percentage of Married Individuals (PercentMarried), Percentage of people aged 18-24 who graduated high school (PctHS18_24), Percentage of people aged 18-24 who have a bachelor’s degree (PctBachDeg18_24), Percentage of people with private health insurance (PctEmpPrivCoverage), and Percentage of people with public health insurance(PctPublicCoverage). These predictors provide valuable insights into the socioeconomic and demographic factors influencing cancer mortality rates. However, predictors such as Average House Size

(AvgHouseholdSize) and percentage of people aged 18-24 who did not graduate high school(PctNoHS18_24) show non-significant coefficients, suggesting they do not impact cancer death rates in this model, possibly due to multicollinearity (High correlation) or other underlying factors that merit further investigation.

The R-squared value of 0.2978 indicates that approximately 29.78% of the variability in cancer death rates is explained by the model's predictors. The adjusted R-squared value suggests that while these predictors account for a significant portion of the variability, other unaccounted-for factors may still influence mortality rates.

```
mod1 <- lm(TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent + AvgHouseholdS
summary(mod1)
```

```
Call:
lm(formula = TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent +
    AvgHouseholdSize + PercentMarried + PctNoHS18_24 + PctHS18_24 +
    PctBachDeg18_24 + PctEmpPrivCoverage + PctPublicCoverage,
    data = filt_project)

Residuals:
     Min       1Q   Median       3Q      Max
-106.182  -13.237    0.419   13.215  160.560

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.062e+02  1.088e+01   9.760  < 2e-16 ***
medIncome          -4.054e-04  7.445e-05  -5.445 5.59e-08 ***
popEst2015         -3.606e-06  1.389e-06  -2.597 0.009443 **
povertyPercent      1.129e+00  1.512e-01   7.468 1.06e-13 ***
AvgHouseholdSize   -1.039e+00  1.057e+00  -0.983 0.325930
PercentMarried     -3.323e-01  9.109e-02  -3.648 0.000269 ***
PctNoHS18_24       -6.850e-02  6.081e-02  -1.127 0.260041
PctHS18_24          6.041e-01  5.290e-02  11.418  < 2e-16 ***
PctBachDeg18_24    -3.601e-01  1.212e-01  -2.971 0.002989 **
PctEmpPrivCoverage  9.979e-01  8.387e-02  11.898  < 2e-16 ***
PctPublicCoverage   9.330e-01  9.711e-02   9.608  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.29 on 3036 degrees of freedom
Multiple R-squared:  0.2978,    Adjusted R-squared:  0.2955
F-statistic: 128.8 on 10 and 3036 DF,  p-value: < 2.2e-16
```

Model Selection Techniques

In our automated model selection process, we utilize two principal techniques: fastbw for fast backward selection and stepAIC for stepwise selection based on the Akaike Information Criterion (AIC). These methods are integral in refining our model by efficiently pinpointing impactful predictors and excluding those that do not significantly explain the variance in cancer death rates, ensuring the model's robustness and interpretability.

Fast Backward Selection (fastbw)

Fastbw starts with all variables and iteratively removes the least significant variable until the best subset of predictors is identified, employing a predefined criterion. Using the fastbw() function, we conducted feature selection to identify the most significant predictors. As a result, variables such as “AvgHouseholdSize” and the “PctNoHS18_24” were deemed non-significant and subsequently removed from the model. The reduced model highlighted predictors such as “medIncome”, “povertyPercent”, and “PctEmpPrivCoverage” and “PctPublicCoverage” as significant factors influencing cancer death rates. Notably, higher median income and lower poverty percentages correlated with lower cancer death rates, while higher percentages of the population with health coverage correlated with lower cancer death rates. The R-squared value of the final model was 0.2976, indicating that the selected predictors explained approximately 29.76% of the variance in cancer death rates.

```
#Fit above model using ols() function in order to use fastbw() function
ols.filt_project <- ols(TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent +

#Perform p-value based selection using fastbw() function
fastbw(ols.filt_project, rule = "p", sls = 0.05)
```

```
 Deleted          Chi-Sq d.f. P      Residual d.f. P      AIC   R2
 AvgHouseholdSize 0.97   1    0.3259 0.97     1    0.3259 -1.03 0.298
 PctNoHS18_24     1.36   1    0.2432 2.33     2    0.3123 -1.67 0.297

Approximate Estimates after Deleting Factors

                       Coef      S.E. Wald Z         P
Intercept           1.026e+02 1.058e+01  9.703 0.000e+00
medIncome          -4.320e-04 7.218e-05 -5.985 2.158e-09
popEst2015         -3.644e-06 1.385e-06 -2.631 8.517e-03
povertyPercent      1.083e+00 1.476e-01  7.338 2.170e-13
PercentMarried     -3.439e-01 9.048e-02 -3.801 1.439e-04
PctHS18_24          6.087e-01 5.236e-02 11.626 0.000e+00
PctBachDeg18_24    -3.104e-01 1.163e-01 -2.670 7.590e-03
PctEmpPrivCoverage  1.030e+00 8.080e-02 12.749 0.000e+00
PctPublicCoverage   9.496e-01 9.558e-02  9.936 0.000e+00

Factors in Final Model

[1] medIncome          popEst2015         povertyPercent     PercentMarried
[5] PctHS18_24         PctBachDeg18_24    PctEmpPrivCoverage PctPublicCoverage
```

```
#Fit reduced model
mod2fbws.filt_project <- lm(TARGET_deathRate ~ medIncome + popEst2015 + povertyPercer


# Print the reduced model
summary(mod2fbws.filt_project)
```

```
Call:
lm(formula = TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent +
```

```
    PercentMarried + PctHS18_24 + PctBachDeg18_24 + PctEmpPrivCoverage +
    PctPublicCoverage, data = filt_project)

Residuals:
     Min       1Q   Median       3Q      Max
-106.813  -13.062    0.609   13.139  160.100

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.026e+02  1.058e+01   9.702  < 2e-16 ***
medIncome          -4.320e-04  7.218e-05  -5.985 2.41e-09 ***
popEst2015         -3.644e-06  1.385e-06  -2.631 0.008564 **
povertyPercent      1.083e+00  1.477e-01   7.337 2.78e-13 ***
PercentMarried     -3.439e-01  9.048e-02  -3.801 0.000147 ***
PctHS18_24          6.087e-01  5.236e-02  11.625  < 2e-16 ***
PctBachDeg18_24    -3.104e-01  1.163e-01  -2.670 0.007633 **
PctEmpPrivCoverage  1.030e+00  8.080e-02  12.748  < 2e-16 ***
PctPublicCoverage   9.496e-01  9.558e-02   9.935  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.29 on 3038 degrees of freedom
Multiple R-squared:  0.2973,    Adjusted R-squared:  0.2954
F-statistic: 160.6 on 8 and 3038 DF,  p-value: < 2.2e-16
```

Stepwise Selection Based on AIC (stepAIC)

StepAIC evaluates models based on the AIC, which balances model fit against complexity by adding or removing predictors iteratively to balance model complexity and performance. Starting with a full model containing all predictor variables, we assessed the inclusion or exclusion of each predictor based on its impact on the AIC value. During this procedure, variables such as “AvgHouseholdSize” and the percentage of people aged “PctNoHS18_24” were removed, similar to the fastbw method. The resulting reduced model exhibited comparable performance in terms of explanatory power, with an R-squared value of approximately 29.73% and a highly significant F-statistic (p-value < 2.2e-16), indicating a good fit for the data. Balancing model complexity and performance, both models offer a refined understanding of the relationship between predictor variables and the target variable while minimizing unnecessary complexity. However, we will continue working with the AIC-reduced model for further analysis.

```
        #####Use stepAIC() function
        aic.mod1 <- stepAIC(mod1)
```

```
Start:  AIC=19195.94
TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent +
    AvgHouseholdSize + PercentMarried + PctNoHS18_24 + PctHS18_24 +
    PctBachDeg18_24 + PctEmpPrivCoverage + PctPublicCoverage

                    Df Sum of Sq     RSS   AIC
- AvgHouseholdSize    1       524 1647809 19195
- PctNoHS18_24        1       689 1647974 19195
<none>                            1647286 19196
- popEst2015          1      3660 1650946 19201
- PctBachDeg18_24     1      4790 1652076 19203
- PercentMarried      1      7220 1654506 19207
```

```
- medIncome           1     16088 1663374 19224
- povertyPercent      1     30260 1677546 19249
- PctPublicCoverage   1     50083 1697369 19285
- PctHS18_24          1     70742 1718027 19322
- PctEmpPrivCoverage  1     76814 1724100 19333

Step:  AIC=19194.91
TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent +
    PercentMarried + PctNoHS18_24 + PctHS18_24 + PctBachDeg18_24 +
    PctEmpPrivCoverage + PctPublicCoverage

                     Df Sum of Sq     RSS   AIC
- PctNoHS18_24        1       739 1648548 19194
<none>                            1647809 19195
- popEst2015          1      3857 1651666 19200
- PctBachDeg18_24     1      4508 1652318 19201
- PercentMarried      1      7195 1655004 19206
- medIncome           1     18072 1665882 19226
- povertyPercent      1     29850 1677659 19248
- PctPublicCoverage   1     53600 1701409 19290
- PctHS18_24          1     70256 1718065 19320
- PctEmpPrivCoverage  1     78631 1726440 19335

Step:  AIC=19194.27
TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent +
    PercentMarried + PctHS18_24 + PctBachDeg18_24 + PctEmpPrivCoverage +
    PctPublicCoverage

                     Df Sum of Sq     RSS   AIC
<none>                            1648548 19194
- popEst2015          1      3755 1652304 19199
- PctBachDeg18_24     1      3867 1652416 19199
- PercentMarried      1      7840 1656389 19207
- medIncome           1     19439 1667987 19228
- povertyPercent      1     29215 1677764 19246
- PctPublicCoverage   1     53561 1702109 19290
- PctHS18_24          1     73335 1721884 19325
- PctEmpPrivCoverage  1     88188 1736736 19351
```

```r
#Fit reduced model
mod2_AIC_rm <- lm(TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent + Percer

summary(mod2_AIC_rm)
```

```
Call:
lm(formula = TARGET_deathRate ~ medIncome + popEst2015 + povertyPercent +
    PercentMarried + PctHS18_24 + PctBachDeg18_24 + PctEmpPrivCoverage +
    PctPublicCoverage, data = filt_project)

Residuals:
     Min       1Q   Median       3Q      Max
-106.813  -13.062    0.609   13.139  160.100
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.026e+02  1.058e+01   9.702  < 2e-16 ***
medIncome          -4.320e-04  7.218e-05  -5.985 2.41e-09 ***
popEst2015         -3.644e-06  1.385e-06  -2.631 0.008564 **
povertyPercent      1.083e+00  1.477e-01   7.337 2.78e-13 ***
PercentMarried     -3.439e-01  9.048e-02  -3.801 0.000147 ***
PctHS18_24          6.087e-01  5.236e-02  11.625  < 2e-16 ***
PctBachDeg18_24    -3.104e-01  1.163e-01  -2.670 0.007633 **
PctEmpPrivCoverage  1.030e+00  8.080e-02  12.748  < 2e-16 ***
PctPublicCoverage   9.496e-01  9.558e-02   9.935  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.29 on 3038 degrees of freedom
Multiple R-squared:  0.2973,    Adjusted R-squared:  0.2954
F-statistic: 160.6 on 8 and 3038 DF,  p-value: < 2.2e-16
```

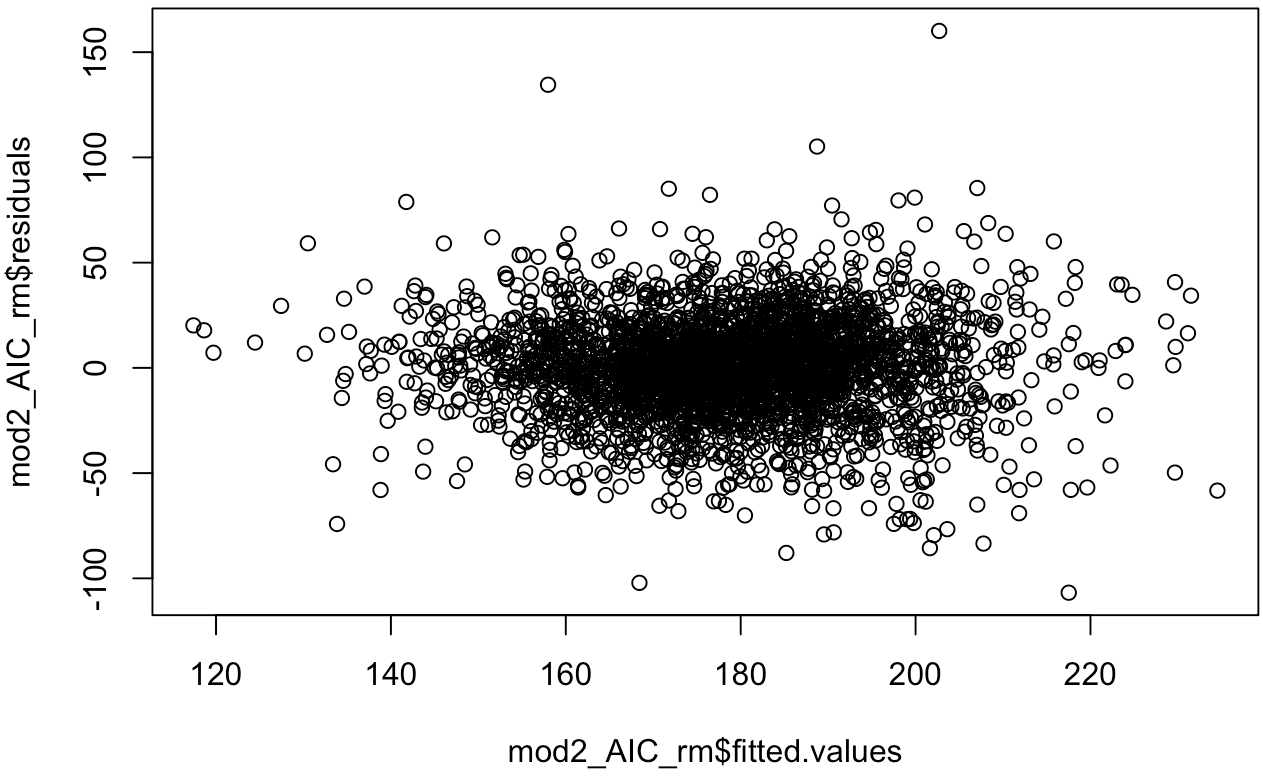
Model Diagnostics

After refining our model, we conducted diagnostic tests to verify the assumptions of linear regression, including linearity, independence, homoscedasticity, and normality of residuals. These diagnostics are crucial as it helps us assess whether the assumptions of linear regression are met and helps identify any potential issues that could affect the model's accuracy and interpretability.

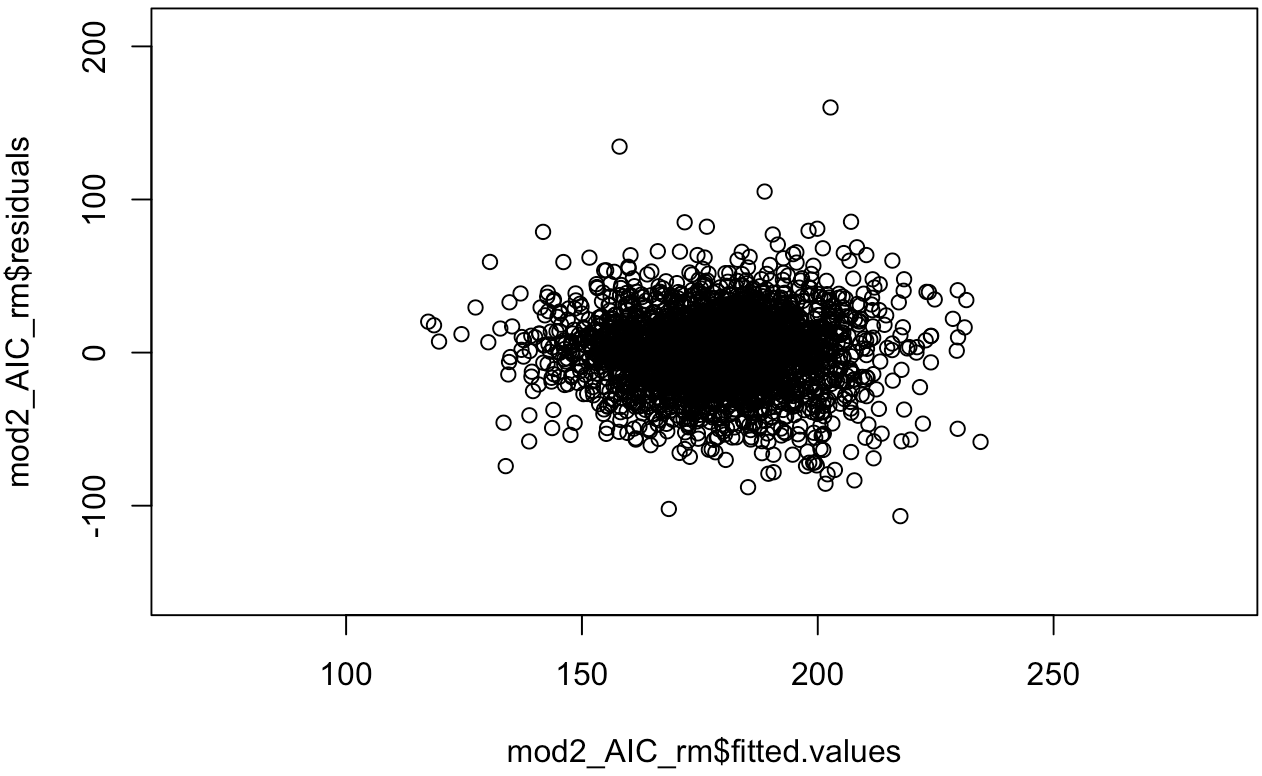
Residual Analysis

Fitted Values vs. Residuals Plot:

```r
#Check assumption of constant error variance
plot(mod2_AIC_rm$fitted.values, mod2_AIC_rm$residuals)
```



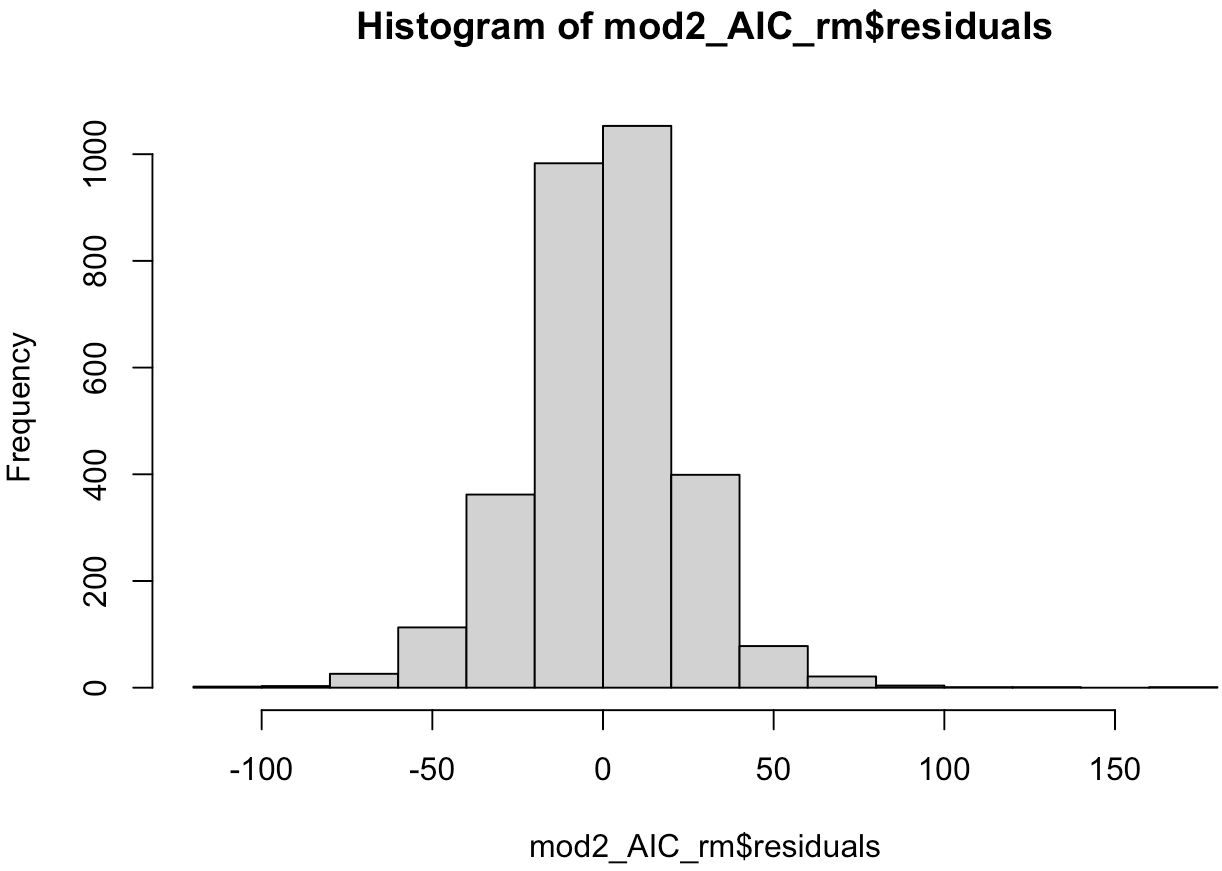

```
# Zoom out to inspect the spread of residuals
plot(mod2_AIC_rm$fitted.values, mod2_AIC_rm$residuals,
     xlim = c(min(mod2_AIC_rm$fitted.values) -50, max(mod2_AIC_rm$fitted.values) + 50
     ylim = c(min(mod2_AIC_rm$residuals) -50, max(mod2_AIC_rm$residuals)+ 50))
```



We examined the relationship between fitted values and residuals to assess homoscedasticity. The random scatter of points around the zero line suggests that the variance of residuals is constant across different levels of fitted values, supporting the assumption of homoscedasticity.

Histogram of Residuals:

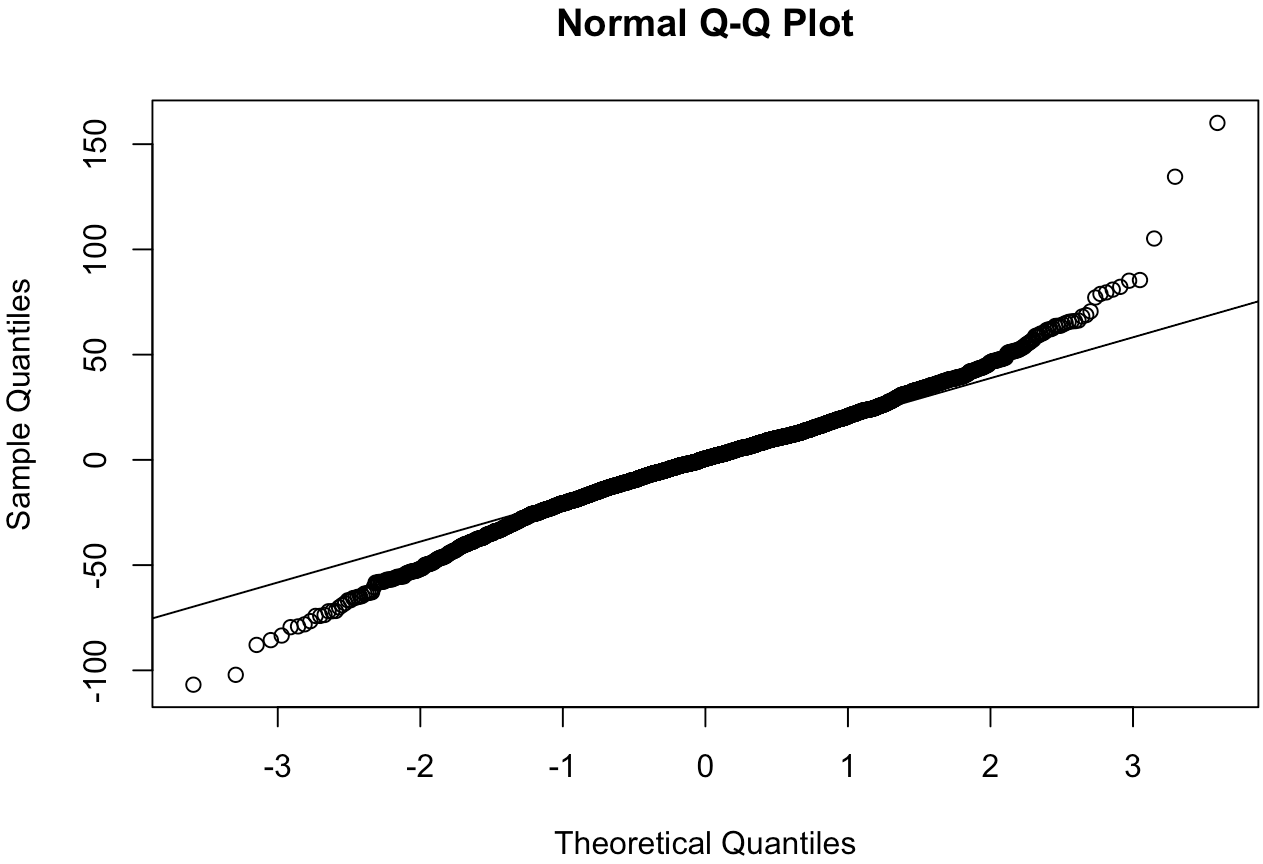
```
# Create a histogram of residuals to assess their distribution
hist(mod2_AIC_rm$residuals)
```



A histogram of residuals showed a roughly symmetric distribution centered around zero, indicating that residuals are approximately normally distributed. This symmetry supports the assumption of normally distributed errors.

Normality Checks: Q-Q Plot and Shapiro-Wilk Test

```
#Check assumption of normal errors
qqnorm(mod2_AIC_rm$residuals)
qqline(residuals(mod2_AIC_rm))
```



The Q-Q plot displayed residuals mostly following the theoretical normal line, with some heavy deviations in the tails.

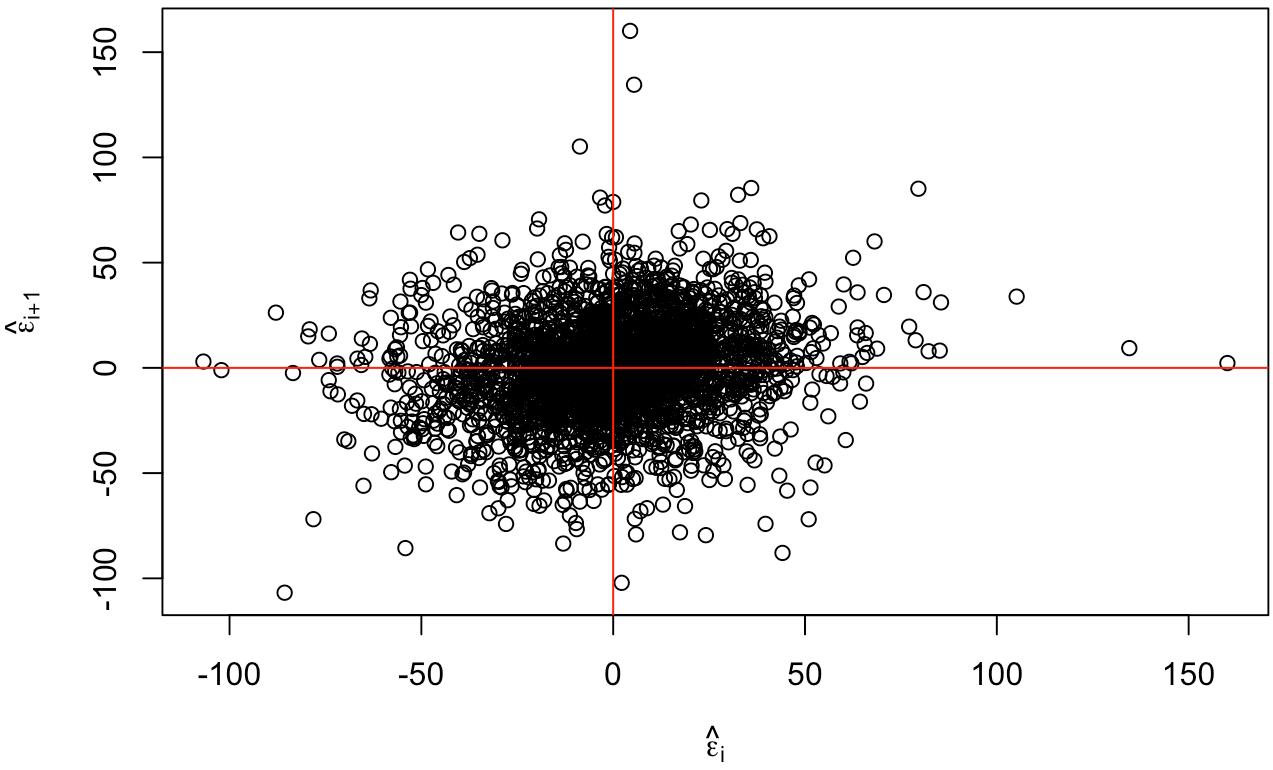
To complement this visual assessment, we conducted the Shapiro-Wilk test, which suggested a slight departure from normality (p-value < 0.05). However, given our large sample size, this test can be sensitive to minor deviations. Thus, we place more weight on the Q-Q plot for assessing normality.

```
        #Check assumption of normal errors
        shapiro.test(mod2_AIC_rm$residuals)
```

```
    Shapiro-Wilk normality test

data:  mod2_AIC_rm$residuals
W = 0.98317, p-value < 2.2e-16
```

Independence Check: Lagged Residual Plot

```
        # plot successive pairs of residuals
        n <- length(residuals(mod2_AIC_rm))
        plot(tail(residuals(mod2_AIC_rm), n-1) ~ head(residuals(mod2_AIC_rm), n-1), xlab = ex
        abline(h = 0,v = 0,col = "red")
```



To assess the independence of residuals, we analyzed a lagged residual plot. The absence of any discernible patterns or trends indicates that residuals are independent, confirming the absence of autocorrelation and supporting the overall model adequacy.

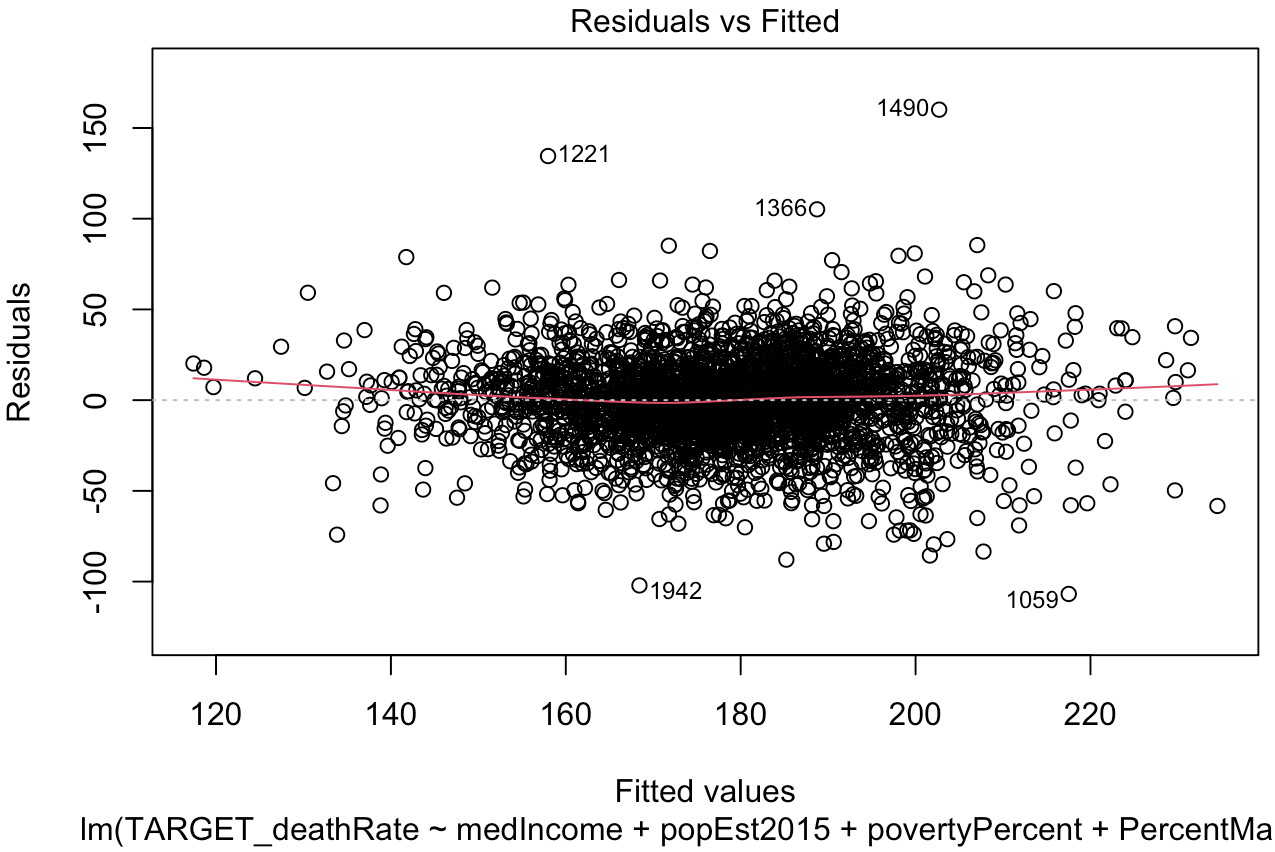
The diagnostic tests confirm that our model meets the key assumptions of linear regression, ensuring its reliability for interpreting the factors influencing cancer death rates. Deviations in normality are noted but are not considered impactful given the model's robustness in other diagnostic areas.

Investigate Fit

After refining our model, we next assess its robustness by investigating the fit for individual observations. This involves identifying outliers and influential points that might unduly affect the model's predictions.

Residual Analysis

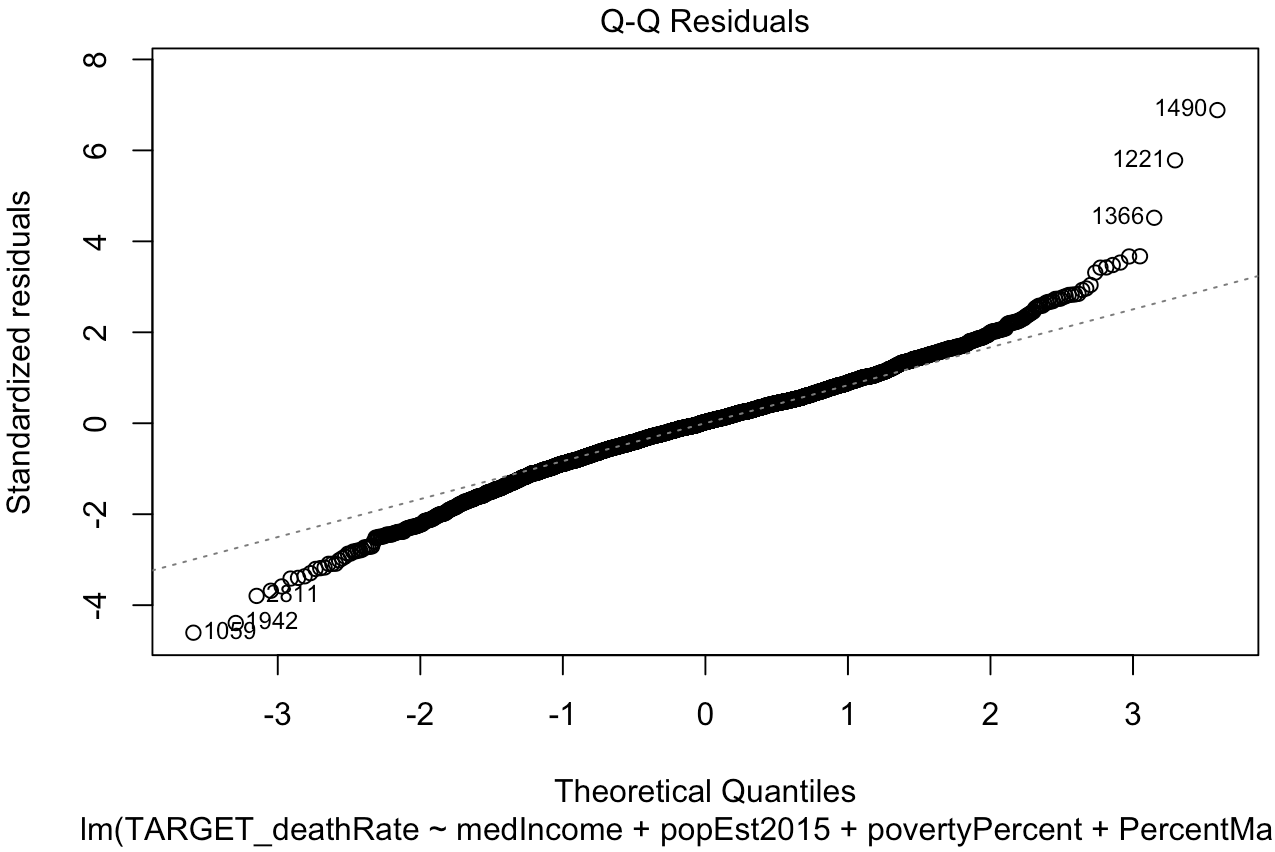
```
        # Residual plot
        plot(mod2_AIC_rm, which = 1, id.n = 5, system = "ggplot2")
```



We begin with a residual plot to detect outliers and influential observations. We identified five data points (with indices 1221, 2727, 1366, 1942, and 1059) that appeared to exert notable influence on the model. These points were labeled using the id.n argument for better identification. The points that deviate significantly from the rest of the data can indicate potential issues.

QQ Plot Analysis

```
# QQ plot
plot(mod2_AIC_rm, which = 2, id.n = 6, system = "ggplot2")
```



Revisiting the QQ plot from earlier, we highlighted the six most extreme points in the plot, identified as data points 1059, 1942, 3811, 2727, 1366, and 1221. As we mentioned earlier, these points exhibited deviations from the expected line, particularly in the distribution's tails. Such deviations may indicate potential outliers or influential observations within the dataset.

Cook's distance

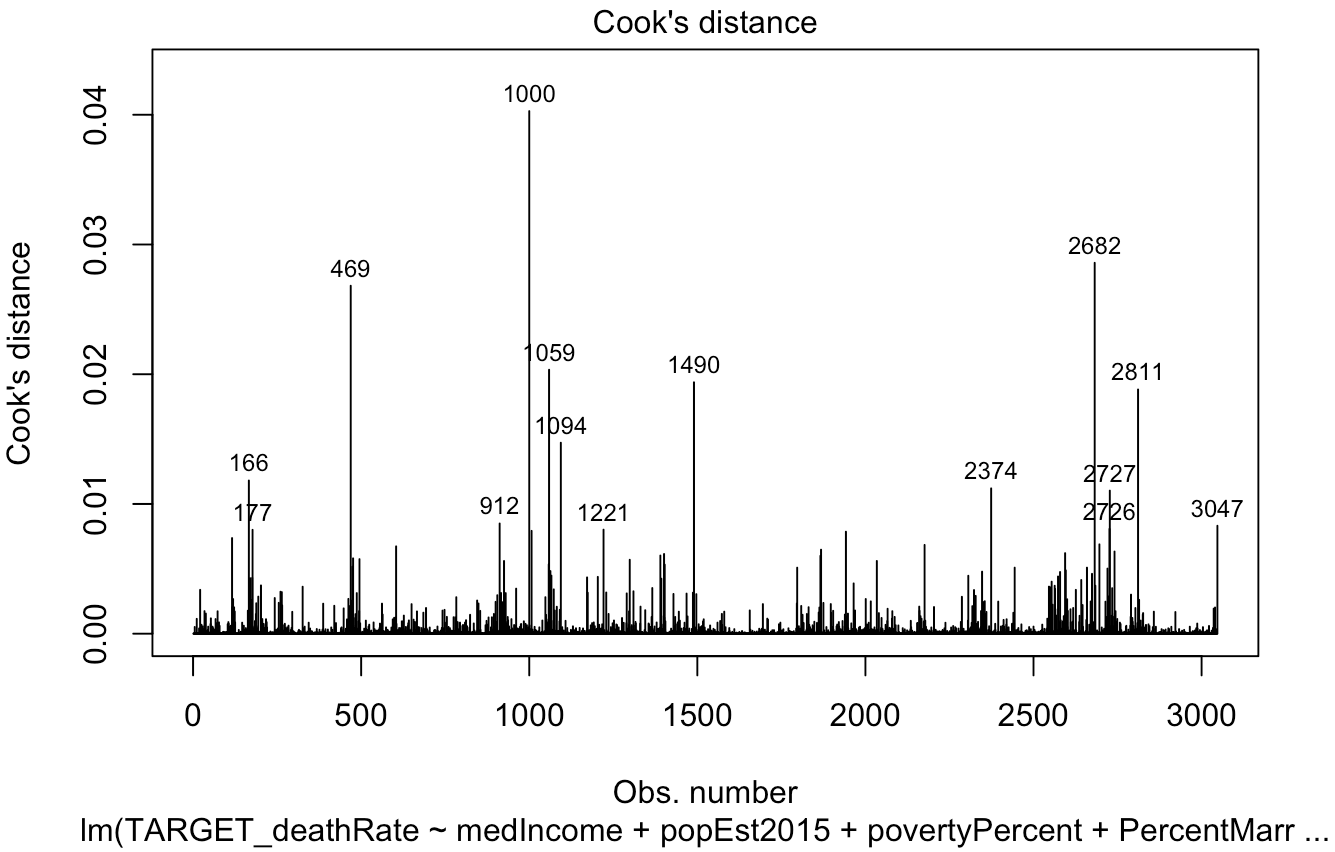
To further investigate our model, we employed Cook's Distance, a measure that assesses the influence of individual data points on both the regression coefficients and the overall model fit. Cook's Distance evaluates how much each observation impacts the fitted values across the entire model, providing crucial insights into which data points might be disproportionately influencing the regression results.

We applied two distinct threshold methods to gauge the influence of each observation. First, we utilized the 50th Percentile of the F-distribution method, This method uses the median influence level as a benchmark, which is derived from the F-distribution. It provides a statistical baseline for what constitutes 'typical' influence within the context of our model. However, this approach did not flag any observations as excessively influential, suggesting that no single observation was beyond the median influence level expected under the model's assumptions.

We also used a more conventional threshold of $\frac{4}{n-k-1}$, where "n" is the number of observations in our dataset, "k" is the number of predictors, and the subtraction of one accounts for the intercept. This method identified 207 observations as potentially influential, which aligns with some of the observations flagged in our other plots. This discrepancy between the two methods highlights the importance of selecting an appropriate threshold based on the model's context and the sensitivity of the analysis to influential data points.

The decision to prioritize the Traditional Threshold method over the 50th Percentile of the F-distribution method was influenced by several factors. We took into account the robustness of our analysis, the potential consequences of excluding influential observations, and the overall objectives of our research. While the 50th Percentile method provides a baseline reference point, the Traditional Threshold method provides a more stringent approach, that helps ensure all potentially problematic observations are scrutinized and addressed appropriately.

```
# Plot Cook's distance
plot(mod2_AIC_rm, which = 4, id.n = 15, system = "ggplot2")
```




```
#Find threshold for Cooks distance, 50th percentile of F distribution with
n <- dim(model.matrix(mod2_AIC_rm))[1]
p <- dim(model.matrix(mod2_AIC_rm))[2]
num.df <- p
den.df <- n-p
F.thresh <- qf(0.50,num.df,den.df)
F.thresh
```

```
[1] 0.9271863
```

```
# > F.thresh
# [1] 0.9271863

#calculate 50th percentile Cook's distances
calculate <- cooks.distance(mod2_AIC_rm)
```

```r
#calculate Cook's 50th percentile distances
which(calculate>F.thresh)
```

```
named integer(0)
```

```r
### Calculate Cook's without 50th percentile
# Identify observations with Cook's distance exceeding threshold
cooks_obs <- cooks.distance(mod2_AIC_rm)
threshold <- 4 / (3047 - 8 - 1)

# Printed as a vector due to the results being duplicated
head(as.vector(which(cooks_obs > threshold)))
```

```
[1]  21  34  38  73 116 117
```

Leverage Analysis

Next we turned our attention to leverage, which measures how much each observation influences the regression coefficients. High leverage points can significantly "pull" the regression line, affecting the model's fit and accuracy.

We began by calculating the threshold for identifying extreme leverage values. Utilizing the rule of thumb formula, $2 * \frac{p}{n}$, where 'p' represents the number of predictors and 'n' is the sample size, we calculated a leverage threshold of 0.007599309. This threshold helps identify observations that might disproportionately influence the model. We identified 224 observations with leverage values surpassing the established threshold.

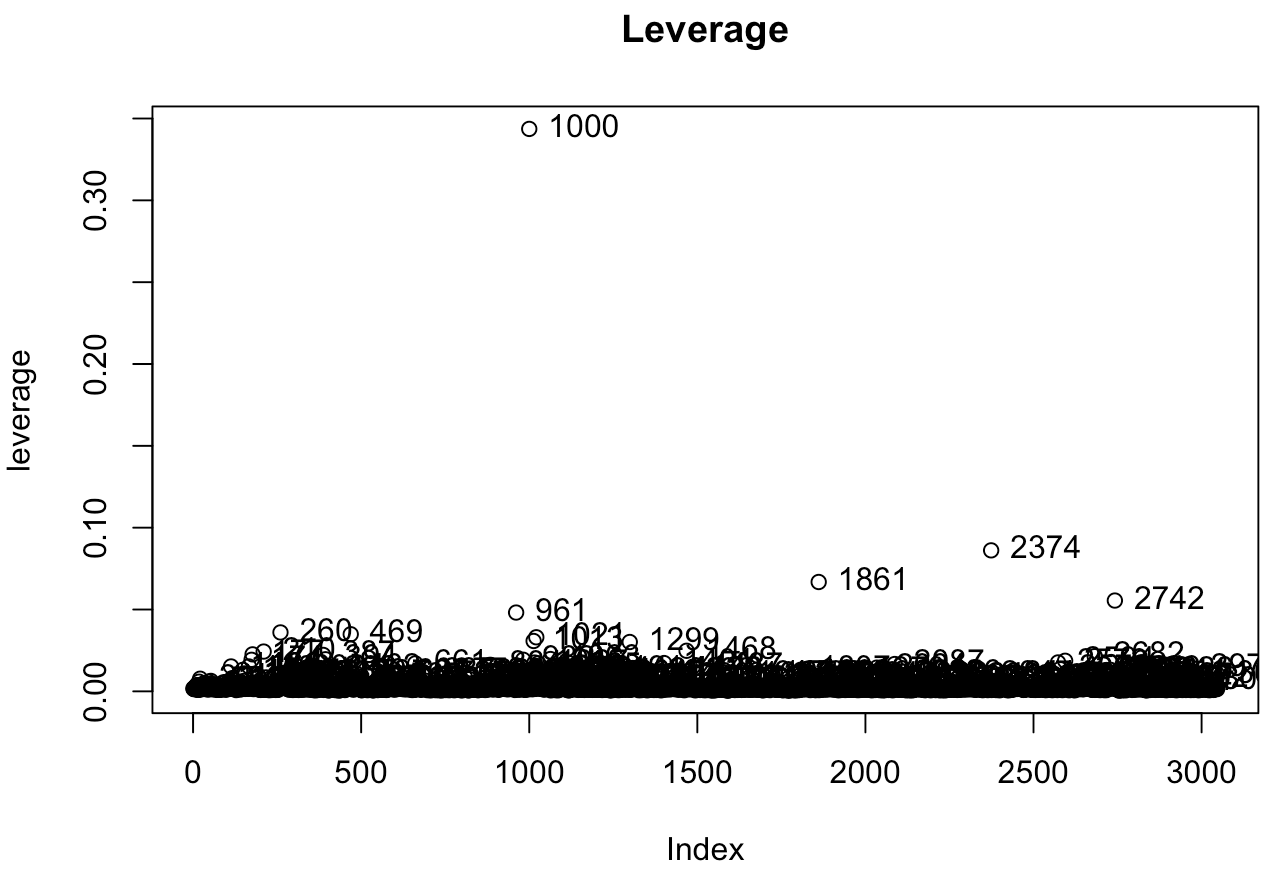
```r
# Calculates the leverage values
leverage <- hatvalues(mod2_AIC_rm)

# Identify the leverage value determining threshold
threshold <- 2 * length(coefficients(mod2_AIC_rm)) / length(residuals(mod2_AIC_rm))

# Identify observations exceeding threshold
extreme_obs <- which(leverage > threshold)
#extreme_obs

# Printed as a vector due to the results being duplicated
head(as.vector(which(leverage > threshold)))
```

```
[1]  21  53 106 113 116 117
```

We generated a plot displaying the leverage values against the observation indices to visualize the leverage values across the dataset. Notable observations, such as 950, 2256, 1766, 2606, 2548, and 913, emerged as influential with exceptionally high leverage highlighted for detailed scrutiny.

```r
# Plot the observations identified
plot(leverage, main = "Leverage", system = "ggplot2")
```

```
# Label influential points on the plot
text(extreme_obs, leverage[extreme_obs], labels = extreme_obs, pos = 4)
```



Standardized Residuals

In our final analysis, we focused on identifying potential outliers and influential observations using standardized residuals. Observations with standardized residuals exceeding three standard deviations from the mean are considered significant deviations and warrant further investigation.

We calculated standardized residuals in two ways to ensure accuracy. First we used the rstandard() function for an automated approach, and then manually, to cross-validate the automated results.

Our analysis revealed 11 observations with standardized residuals greater than +3 or less than -3, indicating significant deviations. Additionally, we identified 54 observations with standardized residuals equal to +3 or -3, positioned at the threshold of being considered outliers.

This method allows us to pinpoint observations that significantly diverge from expected patterns, helping us assess their impact on the regression model's reliability and accuracy.

```
#Compute standardized residuals using rstandard()
stan_ris <- data.frame(round(rstandard(mod2_AIC_rm)))

#moved from end to the top
# Identify outliers on the edge
edge <- stan_ris %>% filter(round.rstandard.mod2_AIC_rm.. == 3 | round.rstandard.mod2
edge
```

	round.rstandard.mod2_AIC_rm..
	<dbl>
34	-3
116	3
122	3
166	3
170	-3
176	-3
326	-3
469	3
472	3
474	-3

1-10 of 54 rows Previous **1** 2 3 4 5 6 Next

```
# Identify outliers,
outliers <- stan_ris %>% filter(round.rstandard.mod2_AIC_rm.. > 3 | round.rstandard.m
outliers
```

	round.rstandard.mod2_AIC_rm..
	<dbl>
1058	-4
1059	-5
1221	6
1366	5
1490	7
1497	4
1942	-4
2598	4
2642	-4
2727	4

1-10 of 11 rows Previous **1** 2 Next

```
############## Manual Calculation to verify #############
#Match rstandard() output "by hand"
n <- dim(model.matrix(mod2_AIC_rm))[1]
p <- dim(model.matrix(mod2_AIC_rm))[2]
RSS <- sum(mod2_AIC_rm$residuals^2)
sigma.hat <- sqrt(RSS/(n-p))

rstandard.match <- mod2_AIC_rm$residuals/(sigma.hat*sqrt(1-hatvalues(mod2_AIC_rm)))
#data.frame(round(rstandard.match))
```

Decision on Handling Outliers

Initially, we considered removing outliers and influential observations from our dataset. However, further analysis showed that these outliers did not significantly affect the model's accuracy or violate key

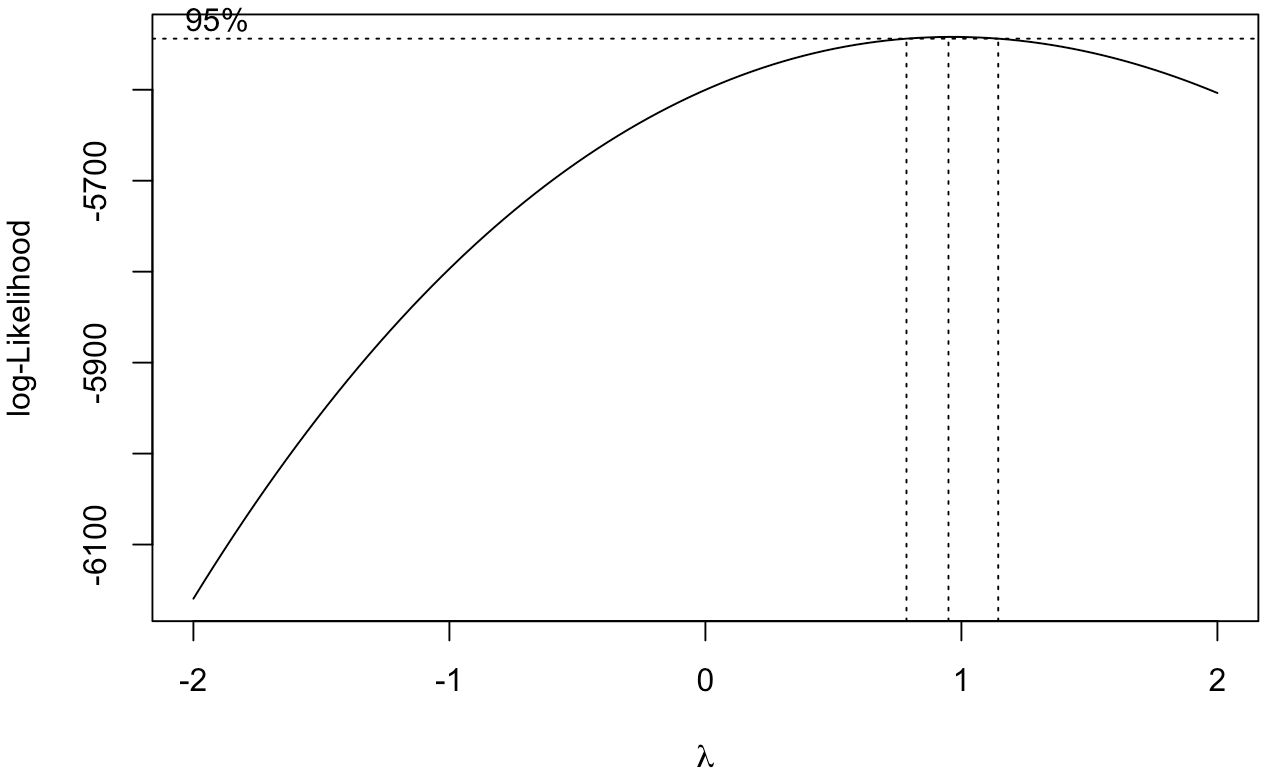
assumptions like homoscedasticity and normality of residuals. Given their minimal impact and the potential loss of valuable data, we decided to retain these points in the dataset. This approach ensures that our model reflects the true dynamics influencing cancer death rates, providing a more comprehensive and accurate analysis.

Apply Transformations

After investigating individual observations and addressing outliers, we identified the need for further modifications to improve our model. Specifically, using the Box-Cox transformation to address potential heteroscedasticity, which will enhance both model accuracy and adherence to linear regression assumptions.

We applied the Box-Cox method to our regression model, identifying an optimal lambda of 0.7878788. This value suggests a moderate transformation, aimed at stabilizing variance across the range of predicted values.

```
#Try Box-Cox to remove heteroscedasticity
bc <- boxcox(mod2_AIC_rm, plotit = T)
```

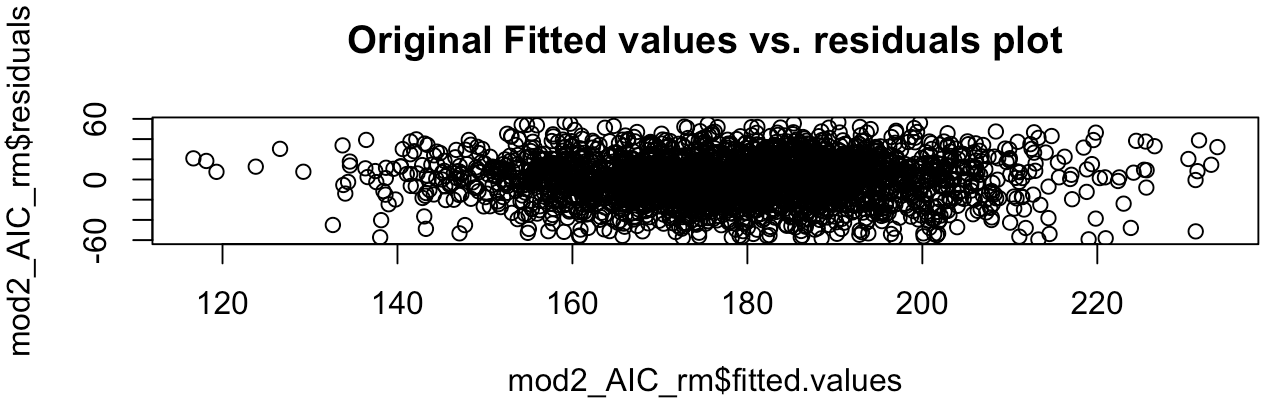


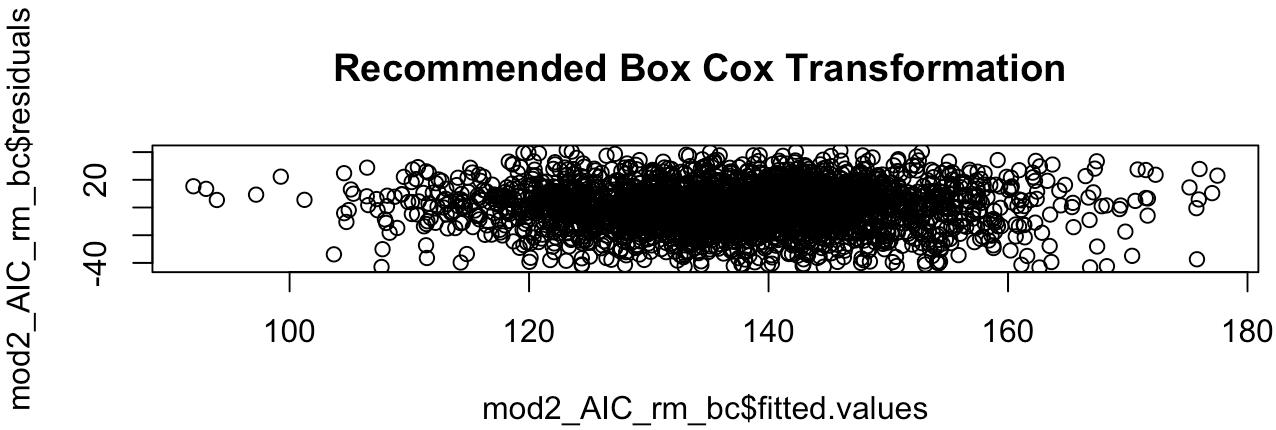

```
# optimal value of lambda
lambda <- bc$x[which.max(bc$y)]
# lambda

# cat("The optimal value of lambda is", lambda)
```

```
# Fit a new model using the transformation recommended
mod2_AIC_rm_bc <- lm(TARGET_deathRate^lambda ~ medIncome + popEst2015 + povertyPercer
sumary(mod2_AIC_rm_bc)
```

```
                    Estimate  Std. Error t value  Pr(>|t|)
(Intercept)         7.9778e+01  7.0832e+00 11.2631 < 2.2e-16
medIncome          -3.1196e-04  4.7602e-05 -6.5534 6.604e-11
popEst2015         -2.5086e-06  9.0558e-07 -2.7702  0.005638
povertyPercent      8.2488e-01  9.9127e-02  8.3214 < 2.2e-16
PercentMarried     -2.5248e-01  6.0983e-02 -4.1402 3.567e-05
PctHS18_24          4.5769e-01  3.4890e-02 13.1180 < 2.2e-16
PctBachDeg18_24    -2.2452e-01  7.8176e-02 -2.8720  0.004108
PctEmpPrivCoverage  7.5655e-01  5.4103e-02 13.9835 < 2.2e-16
PctPublicCoverage   7.1562e-01  6.3989e-02 11.1834 < 2.2e-16

n = 2982, p = 9, Residual SE = 15.18632, R-Squared = 0.35
```

```
# Plotting together to compare
par(mfrow = c(2, 1))
plot(mod2_AIC_rm$fitted.values, mod2_AIC_rm$residuals, main = "Original Fitted values
plot(mod2_AIC_rm_bc$fitted.values, mod2_AIC_rm_bc$residuals, main = "Recommended Box
```





After fitting a new model using the recommended Box-Cox transformation, we evaluated the model's performance by comparing our original diagnostic plot with our new recommended transformation. The diagnostic plots did not visually confirm a significant change. This could be due to the scale or sensitivity of the plot used.

We then assessed our models' fit by examining the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. The AIC and BIC values serve as model performance indicators, with lower values suggesting a better fit. Our comparison found that the model with the recommended lambda (0.7878788) had substantially lower AIC (19693.76) and BIC values (19753.98) than the original model (27843.28 and 27903.5). Despite the lack of discernible differences in the plots, the lower AIC and BIC values indicate that the Box-Cox transformation improved the model's fit and addressed heteroscedasticity.

```
# Calculate AIC and BIC for each model
aic_Orig_mod <- AIC(mod2_AIC_rm)
bic_Orig_mod <- BIC(mod2_AIC_rm)

aic_Recom_mod <- AIC(mod2_AIC_rm_bc)
bic_Recom_mod <- BIC(mod2_AIC_rm_bc)

# Display AIC and BIC values
cat("The AIC for the original model is,", aic_Orig_mod, "and the BIC values is,",bic_
```

```
The AIC for the original model is, 26565.58 and the BIC values is, 26625.58
```

```
cat("The AIC for the recommended lambda model is,", aic_Recom_mod, "and the BIC value
```

```
The AIC for the recommended lambda model is, 24697.97 and the BIC values is, 24757.98
```

Final Model Summary and Predictions

After addressing heteroscedasticity the coefficients from our final regression model, which includes a Box-Cox transformation, provide insights into how various socioeconomic factors influence the Transformed Target Death Rate. Each coefficient represents the expected change in the death rate for a one-unit change in the predictor, holding all other variables constant. The model's intercept, is approximately 39.82865, representing the expected Transformed Target Death Rate when all predictors are zero. This value serves as a baseline against which the effects of other variables are measured.

```
# Extract coefficient estimates and p-values
coefficients <- coef(summary(mod2_AIC_rm_bc))
parameter_estimates <- coefficients[, "Estimate"]
p_values <- coefficients[, "Pr(>|t|)"]


# Combine into a data frame to match example
bc_final_model_summary <- data.frame(
  Predictor = rownames(coefficients),
  Parameter_Estimate = parameter_estimates,
  p_value = p_values,
  row.names = NULL
)
bc_final_model_summary
```

Predictor	Parameter_Estimate	p_value
<chr>	<dbl>	<dbl>
(Intercept)	7.977803e+01	7.601237e-29

Predictor	Parameter_Estimate	p_value
<chr>	<dbl>	<dbl>
medIncome	-3.119591e-04	6.603508e-11
popEst2015	-2.508596e-06	5.637796e-03
povertyPercent	8.248784e-01	1.307746e-16
PercentMarried	-2.524782e-01	3.567160e-05
PctHS18_24	4.576879e-01	2.939287e-38
PctBachDeg18_24	-2.245211e-01	4.107855e-03
PctEmpPrivCoverage	7.565454e-01	4.421451e-43
PctPublicCoverage	7.156172e-01	1.803504e-28

9 rows

The “medIncome” coefficient of approximately -1.182609e-04 indicates that an increase in median income is associated with a slight decrease in the Transformed Target Death Rate. This suggests that higher income levels may contribute to lower cancer death rates, possibly due to better access to healthcare or healthier living conditions. Likewise, Population Estimate 2015 (popEst2015) has a coefficient of approximately -9.223653e-07, this suggests that larger populations slightly decrease the Transformed Target Death Rate. This could reflect the effects of urbanization or better healthcare infrastructure in more populous areas. “povertyPercent” had a positive coefficient of 2.699666e-01 which implies that higher poverty rates are associated with higher cancer death rates, underscoring the negative impact of poverty on public health. Looking at the “PercentMarried”, the negative coefficient of -9.519539e-02 indicates that higher marriage rates are associated with lower cancer death rates, possibly due to factors like increased social support and shared health benefits. The remaining coefficients for educational attainment, health coverage, and other factors follow a similar pattern, and a distinct impact on the Transformed Target Death Rate.

```
summary(mod2_AIC_rm_bc)
```

```
Call:
lm(formula = TARGET_deathRate^lambda ~ medIncome + popEst2015 +
    povertyPercent + PercentMarried + PctHS18_24 + PctBachDeg18_24 +
    PctEmpPrivCoverage + PctPublicCoverage, data = filt_project)

Residuals:
    Min      1Q  Median      3Q     Max
-43.292  -9.247   0.415   9.423  41.430

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.978e+01  7.083e+00  11.263  < 2e-16 ***
medIncome          -3.120e-04  4.760e-05  -6.553 6.60e-11 ***
popEst2015         -2.509e-06  9.056e-07  -2.770  0.00564 **
povertyPercent      8.249e-01  9.913e-02   8.321  < 2e-16 ***
PercentMarried     -2.525e-01  6.098e-02  -4.140 3.57e-05 ***
PctHS18_24          4.577e-01  3.489e-02  13.118  < 2e-16 ***
PctBachDeg18_24    -2.245e-01  7.818e-02  -2.872  0.00411 **
PctEmpPrivCoverage  7.565e-01  5.410e-02  13.983  < 2e-16 ***
PctPublicCoverage   7.156e-01  6.399e-02  11.183  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 15.19 on 2973 degrees of freedom
Multiple R-squared:  0.353, Adjusted R-squared:  0.3512
F-statistic: 202.7 on 8 and 2973 DF,  p-value: < 2.2e-16
```

```
        # cat("The R2 for the Box Cox model is:", round(summary(mod2_AIC_rm_bc)$r.squared,2))
```

The summary of our linear regression model, incorporating the Box-Cox transformation, provides crucial insights into the model's effectiveness and the relationships between predictors and the Transformed Target Death Rate.

Firstly, the transformed predictor variables exhibit statistically significant relationships with the Transformed Target Death Rate, and the model's overall performance is assessed using the multiple R-squared value of 0.2961 which indicates that approximately 29.61% of the variability in the Transformed Target Death Rate is explained by the predictors included in the model. This suggests a moderate level of explanatory power, highlighting the influence of the included variables on the outcome. We see the adjusted R-squared at 0.2942, this adjusted metric accounts for the number of predictors, offering a slightly conservative estimate compared to the R-squared. It confirms that the model remains effective even after adjusting for the complexity added by multiple predictors. The F-statistic of 159.7 with a p-value less than 2.2e-16 robustly indicates that the model is statistically significant. This confirms that the predictors, collectively, have a non-zero effect on the Transformed Target Death Rate, validating the model's predictive relevance.

Confidence Intervals in Model Analysis

To quantify the uncertainty associated with our model's estimates, particularly for the median income predictor, we compute 95% confidence intervals. These intervals are crucial as they provide a range within which we can expect the true parameter values to fall 95% of the time, assuming the model is accurate and the data are representative. A narrower confidence interval indicates a higher precision of the estimate, suggesting that we have a more precise understanding of how a predictor affects the target variable. Conversely, a wider interval indicates greater uncertainty. It's important to note that these intervals are calculated on the transformed scale due to the Box-Cox transformation applied to the model. This means the intervals might not directly reflect the scale of the original median income values.

The 95% confidence interval for the transformed median income coefficient ranges from approximately -0.0001554194 to -0.00008110227. This interval suggests that with 95% confidence, the true effect of a one-unit increase in median income on the transformed target death rate lies within this range. The negative values indicate that higher median income is associated with a decrease in the transformed target death rate, reinforcing the importance of economic factors in public health outcomes.

We also computed 95% confidence intervals for all other predictors in the model. Notably, median income and the population estimate for 2015 show relatively narrower intervals compared to other variables, indicating more reliable predictions when using these predictors.

```
        # Compute 95% confidence interval for my choice of most important
        confint(mod2_AIC_rm_bc, "medIncome")
```

```
                2.5 %        97.5 %
medIncome -0.0004052961 -0.0002186221
```

```
        # Calculate all of the predictors
```

```
confint(mod2_AIC_rm_bc) # 95% CIs
```

```
                          2.5 %        97.5 %
(Intercept)         6.588965e+01  9.366641e+01
medIncome          -4.052961e-04 -2.186221e-04
popEst2015         -4.284224e-06 -7.329688e-07
povertyPercent      6.305136e-01  1.019243e+00
PercentMarried     -3.720508e-01 -1.329056e-01
PctHS18_24          3.892766e-01  5.260992e-01
PctBachDeg18_24    -3.778060e-01 -7.123621e-02
PctEmpPrivCoverage  6.504625e-01  8.626283e-01
PctPublicCoverage   5.901500e-01  8.410844e-01
```

To ensure that our findings are practical and easily understandable, we recalculated the 95% confidence intervals using the regression model before applying the Box-Cox transformation. This approach allows stakeholders to directly grasp the impact of socioeconomic changes, such as variations in median income, on cancer death rates without the need to interpret transformed scales.

This interval indicates that we can be 95% confident that the true population effect of a one dollar increase in “medIncome” on the target variable Target Death Rate falls within this range.

```
############## Original Model #############

# Compute 95% confidence interval for my choice of most important
confint(mod2_AIC_rm, "medIncome")
```

```
                2.5 %        97.5 %
medIncome -0.0005505854 -0.0002952674
```

```
# I wanted to look at all of them
confint(mod2_AIC_rm) # 95% CIs
```

```
                          2.5 %        97.5 %
(Intercept)         8.035621e+01  1.183471e+02
medIncome          -5.505854e-04 -2.952674e-04
popEst2015         -5.880877e-06 -1.023750e-06
povertyPercent      8.745197e-01  1.406194e+00
PercentMarried     -5.040876e-01 -1.770036e-01
PctHS18_24          5.333709e-01  7.205060e-01
PctBachDeg18_24    -5.134186e-01 -9.411650e-02
PctEmpPrivCoverage  8.863417e-01  1.176525e+00
PctPublicCoverage   8.049205e-01  1.148129e+00
```

To gain insights into the typical behavior of our model, we calculated the median values of all predictor variables. This approach allows us to estimate the expected Target Death Rate under average conditions, providing a practical perspective on how typical values of predictors relate to the outcome.

The estimated Target Death Rate is approximately 178.465 deaths per 100,000 individuals. This prediction serves as a baseline for evaluating the model’s performance under typical conditions. The confidence interval for this prediction ranges from approximately 177.52 to 179.41 deaths per 100,000 individuals. This interval provides a statistical range that is likely to contain the true Target Death Rate 95% of the time, under the assumption that the model is correct and the data are representative.

```r
# Calculate the median values of predictor variables
mediandata <- data.frame(
  medIncome = median(project$medIncome),
  popEst2015 = median(project$popEst2015),
  povertyPercent = median(project$povertyPercent),
  PercentMarried = median(project$PercentMarried),
  PctHS18_24 = median(project$PctHS18_24),
  PctBachDeg18_24 = median(project$PctBachDeg18_24),
  PctEmpPrivCoverage = median(project$PctEmpPrivCoverage),
  PctPublicCoverage = median(project$PctPublicCoverage))

# Calculate the predicted values using pre bc
predict(mod2_AIC_rm, mediandata, interval = "confidence")
```

```
       fit      lwr      upr
1 178.3812 177.5272 179.2352
```

In this part of our analysis, we focused on estimating the Target Death Rate for Ada County, ID, using average values of key socioeconomic and demographic predictors found on the US Census website. This localized approach helps us understand how the model performs under conditions specific to Boise and surrounding cities, providing insights into the expected cancer death rate in this particular area.

The model predicts a Target Death Rate of approximately 147.49 per 100,000 individuals for Ada County, based on average predictor values reflective of this region. This estimate assumes that the conditions in Ada County align closely with these average values.

Additionally, the model provides a 95% prediction interval ranging from about 101.69 to 193.30 deaths per 100,000 individuals. This wide range suggests significant variability in the data, reflecting the complex interplay of various factors that influence cancer death rates. This wide range highlights the inherent uncertainty and variability in predicting specific instances of the Target Death Rate, acknowledging the influence of both observed and unobserved factors.

```r
############# Original Model #############
# Calculate the predicted values using estimated aveages in Ada County, ID
predict(mod2_AIC_rm, new = data.frame(medIncome = 87774, popEst2015 = 494967, poverty
```

```
       fit      lwr     upr
1 146.9531 106.1052 187.801
```

Conclusion: Summary of Findings and Implications

As we conclude this analysis, we reflect on the significant insights gained from our investigation into the socioeconomic determinants of cancer death rates. This study has not only highlighted the intricate relationships between various economic and demographic factors and health outcomes but also underscored the critical role of data-driven approaches in public health policy and intervention planning.

Key Findings:

1. Impact of Socioeconomic Factors: Our analysis has demonstrated significant relationships between various socioeconomic factors and cancer death rates. Notably, median income and poverty percentage

have shown substantial effects, indicating that economic conditions play a crucial role in public health outcomes.
2. Model Performance: The final regression model, enhanced with a Box-Cox transformation, explains approximately 29.61% of the variability in the transformed cancer death rates. This level of explanatory power, while moderate, provides valuable insights into the factors influencing cancer mortality.
3. Localized Predictions: Applying the model to specific geographic settings, such as Ada County, ID, has allowed us to estimate local cancer death rates. These predictions are crucial for tailoring public health interventions to specific community needs.

Implications:

1. Public Health Strategies: The findings underscore the importance of addressing socioeconomic disparities as part of cancer prevention and control strategies. Policies aimed at improving economic conditions could indirectly contribute to lower cancer mortality rates.
2. Resource Allocation: The model can help public health officials allocate resources more effectively by identifying areas where interventions might have the greatest impact, based on the socioeconomic profiles.
3. Further Research: While the model provides significant insights, the unexplained variability suggests that other factors, possibly environmental or genetic, also influence cancer death rates. Further research is needed to explore these dimensions.

Recommendations:

1. Integrate Socioeconomic Interventions: Encourage the integration of socioeconomic interventions into public health programs aimed at reducing cancer mortality. This could include programs to boost economic development in low-income areas or improve access to education and healthcare.
2. Tailored Public Health Policies: Use the model's predictions to develop tailored public health policies for different regions. Understanding the specific needs and challenges of each area can lead to more effective and targeted interventions.
3. Continuous Model Refinement: Continue to refine the model by incorporating additional data, such as environmental or genetic factors, to improve its accuracy and comprehensiveness. Regular updates and validations of the model with new data are essential to keep the predictions relevant and useful.
4. Stakeholder Engagement: Engage with local communities and stakeholders to ensure that the model's findings are translated into actionable strategies that are culturally appropriate and supported by those they are meant to serve.

This project has not only highlighted the significant impact of socioeconomic factors on cancer death rates but also demonstrated the utility of statistical modeling in public health. By translating complex data into actionable insights, we can inform better public health decisions, ultimately leading to improved health outcomes and reduced disparities.