Demographic Analysis and Mortality Patterns in California Counties A Linear Regression Approach

Mingi Kang

Introduction

In the state of California, there are 58 different counties with a combined state population of 38,940,231 residents (in year 2023). During 2021, there were a total of 334,555 deaths counting all the causes of death. There are different types of demographic of the population that are more likely to get a certain disease or health problem and can die from it. For example, Koreans are more likely to get malignant neoplasms, heart diseases, pneumonia and eventually cause death from these. Although Korean is not a demographic, many of the different demographic groups can have a similar association with various causes of deaths. The relationship between demographic groups and causes of death can be looked into to find different patterns for different kinds of demographic groups. In this project, we consider the relationship of 12 demographic groups with 15 cause of death variable over the 58 counties in California. The main question of the analysis is :

- How do various demographic factors impact different causes of death?

Data source, cleaning, and limitations

The data for the number of deaths in each county in California was downloaded from the California Department of Public Health website, and the page is titled Death Profiles by County. The data folder contained multiple years of final deaths by year, month by occurrence, residence county. The death data that was used in this project was from 2019-2021 Final Deaths by Month by Occurrence County csv file. The data compromised of 13 columns with year, month, county, geography type, strata, strata name, cause, cause description, ICD revision, count, annotation code, annotation description and data revision date. Since the number of deaths were divided into 12 months, we had to do some data cleaning. I have combined the total number of deaths in each month of 2021 by the different cause description and by the county. The causes of death we considered for the analysis were:

- All causes, Alzheimer's disease, Malignant Neoplasms, Chronic lower respiratory diseases, Diabetes Mellitus, Assault (homicide), Heart disease, Essential Hypertension and Hyptertensive Renal disease, Accidents, Chronic Liver disease Cirrhosis, Nephritis Nephrotic Syndrome Nephrosis, Parkinson's disease, Influenza - Pneumonia, Cerebrovascular disease, Intentional self harm - suicide.

The data for the demographic group population in California was downloaded from the California State Association of Counties and the page is titled DataPile. The data we chose is named Current DataPile which is comprised of 4 excel books and the 2023 California counties demographic population data is in the people excel book. From this excel file, there were the population of each counties for different races, ages, registered voters, labor force, poverty. For this analysis, we only used the race and ages population columns. The race demographics that we considered were:

- Total, American Indian, Asian, Black, Hispanic, Multi-Racial/Ethnic, Hawaiian/Pacific Island, and White

The age demographics that we considered were

- ages 0-5, ages 6-17, ages 18-64, ages 65+ population

These two data files were combined into one named demographic_death.csv with the first column consisting of the counties in California and all of the other columns being the race demographic population, age demographic population, number of deaths by different causes.

The full procedure for data cleaning is provided in the Appendix. This process was done in Python and Jupyter Notebook.

Limitations of the data

The summary statistics themselves

The population of the counties in California are very contrasting. With a couple of counties being very small in population, and couple extremely high in population. With this in mind, the average of the population/death numbers does not accurately show the populations. The variance and standard deviation of the population/death will also be high as the outliers of both low and high population counties will force it to be high. The higher population counties will have a higher number of deaths and the lower population counties will have a lower number of deaths.

The considerable disparities in population sizes among the California counties will influence the overall trends in our data as larger counties will exhibit a higher number of deaths.

Exploratory Data Analysis

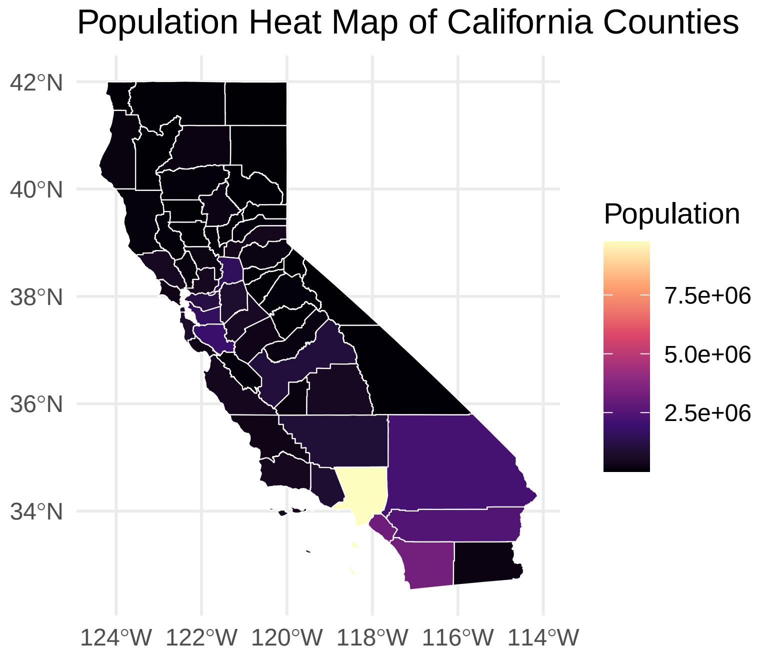
We will examine the population of different demographic groups.

```
data.set$X65_plus
      0.003609701


      X0_5
-0.0117568



  |
  |                                                                      |   0%
  |
  |                                                                      |   1%
  |
  |=                                                                     |   1%
  |
  |=                                                                     |   2%
  |
  |==                                                                    |   2%
  |
  |==                                                                    |   3%
  |
  |===                                                                   |   4%
  |
  |====                                                                  |   5%
  |
  |====                                                                  |   6%
  |
  |=====                                                                 |   7%
  |
  |=====                                                                 |   8%
  |
  |======                                                                |   8%
  |
  |======                                                                |   9%
  |
  |=======                                                               |  10%
  |
  |========                                                              |  11%
  |
|========                                                              |  12%
|
|=========                                                             |  12%
|
|=========                                                             |  13%
|
|==========                                                            |  14%
|
|==========                                                            |  15%
|
|===========                                                           |  15%
|
|===========                                                           |  16%
|
|===============                                                       |  22%
|
|================                                                      |  22%
|
|================                                                      |  23%
|
|=================                                                     |  24%
|
|=================                                                     |  25%
|
|==================                                                    |  25%
|
|==================                                                    |  26%
|
|===================                                                   |  27%
|
|===================                                                   |  28%
|
|====================                                                  |  28%
|
|====================                                                  |  29%
|
|=====================                                                 |  29%
|
|=====================                                                 |  30%
|
|===============================                                       |  44%
|
|===============================                                       |  45%
```

```
|
|================================                                      |  46%
|
|=================================                                     |  47%
|
|=================================                                     |  48%
|
|==================================                                    |  48%
|
|==================================                                    |  49%
|
|===================================                                   |  49%
|
|===================================                                   |  50%
|
|====================================                                  |  51%
|
|====================================                                  |  52%
|
|=====================================                                 |  52%
|
|=====================================                                 |  53%
|
|======================================                                |  54%
|
|======================================                                |  55%
|
|=======================================                               |  55%
|
|=======================================                               |  56%
|
|========================================                              |  57%
|
|========================================                              |  58%
|
|=========================================                             |  58%
|
|=========================================                             |  59%
|
|==========================================                            |  60%
|
|==========================================                            |  61%
|
|===========================================                           |  61%
|
|===========================================                           |  62%
|
|============================================                          |  62%
|
|============================================                          |  63%
|
|=============================================                         |  64%
|
|==============================================                        |  65%
|
|==============================================                        |  66%
|
|===============================================                       |  67%
|
|===============================================                       |  68%
|
|================================================                      |  68%
|
|================================================                      |  69%
|
|=================================================                     |  70%
|
|==================================================                    |  71%
|
|==================================================                    |  72%
|
|===================================================                   |  72%
|
|===================================================                   |  73%
|
|====================================================                  |  74%
|
|=====================================================                 |  75%
|
|=====================================================                 |  76%
|
|======================================================                |  77%
|
|======================================================                |  78%
|
|=======================================================               |  78%
```

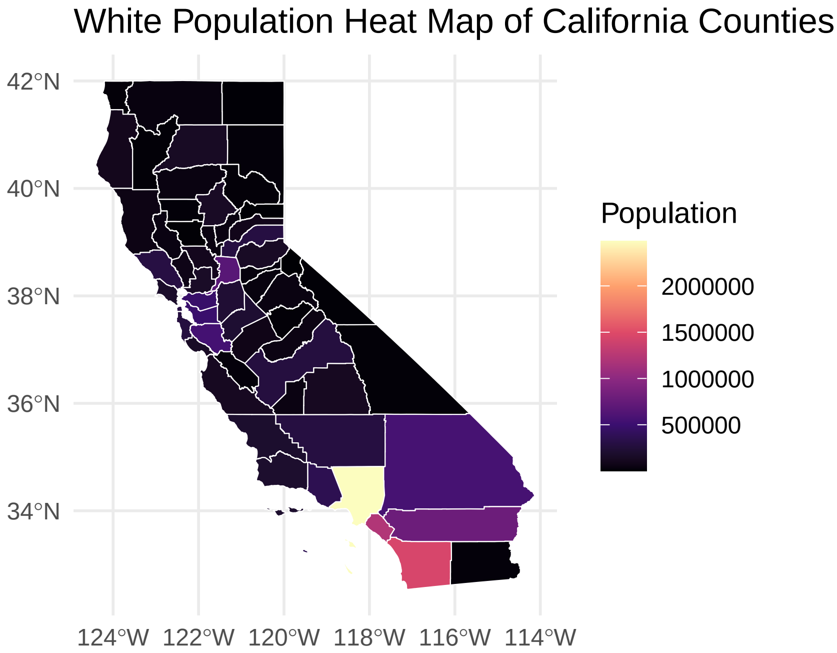
```
  |
  |=======================================================               |  79%
  |
  |========================================================              |  80%
  |
  |=========================================================             |  81%
  |
  |=========================================================             |  82%
  |
  |==========================================================            |  82%
  |
  |==========================================================            |  83%
  |
  |===========================================================           |  84%
  |
  |===========================================================           |  85%
  |
  |============================================================          |  85%
  |
  |============================================================          |  86%
  |
  |=============================================================         |  87%
  |
  |==============================================================        |  88%
  |
  |==============================================================        |  89%
  |
  |===============================================================       |  90%
  |
  |================================================================      |  91%
  |
  |=================================================================     |  92%
  |
  |=================================================================     |  94%
  |
  |======================================================================| 100%
```

```
Retrieving data for the year 2022
```



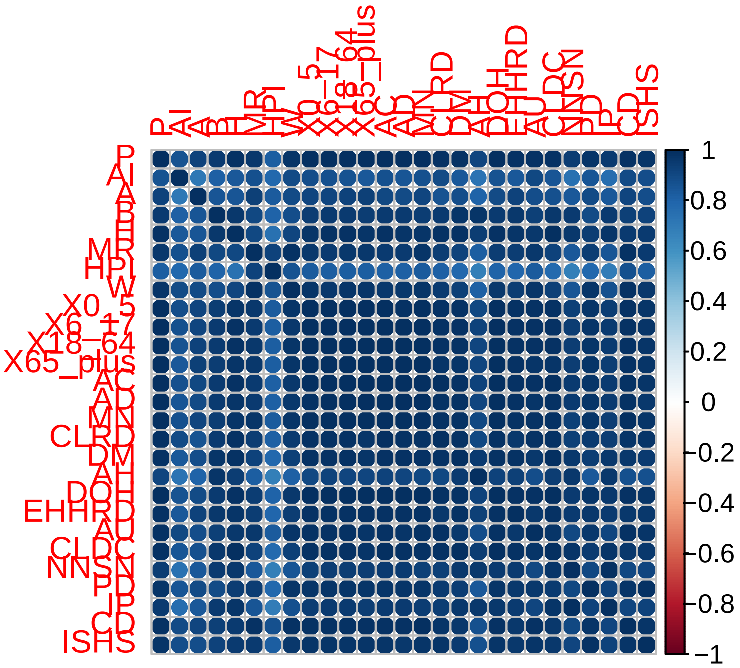
Population Heat Map of California Counties

```
Retrieving data for the year 2022

Retrieving data for the year 2022
```



White Population Heat Map of California Counties

The three Heat Maps show the total population, number of deaths caused by Alzheimer's Disease, and the white population. All three of the plots have similar shading as the different variables whether it is demographic or death, it will be similar as higher populated counties will have larger number of deaths and



The correlation matrix does not do a good job of showing the correlations between the different demographic variables and the cause of death variables. It is hard to see as almost majority of the correlations are blue, indicating that all demographic and death variables are all highly correlated with each other. This does not allow us to figure out which variables will be significant for model selection, and another method will be needed to figure out the best combination of demographic variables for model selection.

Linear Model and Model Selection

The model we will be using for the analysis is the Linear Model with multiple regressors. The data for demographic population and causes of deaths are all numbers with 58 rows of data. We are regressing 12 different demographic variables with each of the 15 death variables. Since the correlation matrix does not help much with viewing which variables will be the best variables, we needed a way to do the garden of forking paths with all 12 demographic variables. If we wanted to try all possibilities of 12 variables, there are 4095 different subsets that can be formed. We will never be able to see which model with a particular subset is the best doing this by hand.

To measure the best fit and best model for the Linear Model, we can use Log Likelihood and AIC. The log ikelihood is a measure to quantify how well the model explains the observed data (number of deaths by cause). The log likelihood is calculated based on the residual sum of squares and the variance of the residuals. The AIC is short for Akaike Information Criterion. AIC shows the model fit and complexity. Higher log likelihood values indicate a better fit of the model to the data. Lower values of AIC indicate better models when comparing across different models.

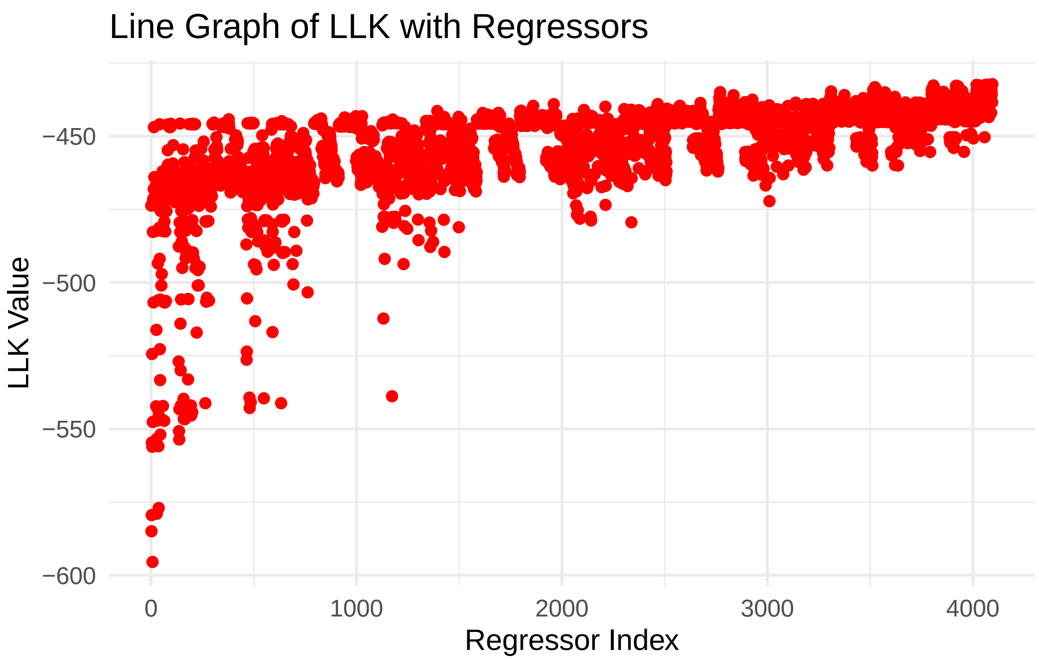
The function for generating all possible linear models of the possible subset of variables is:

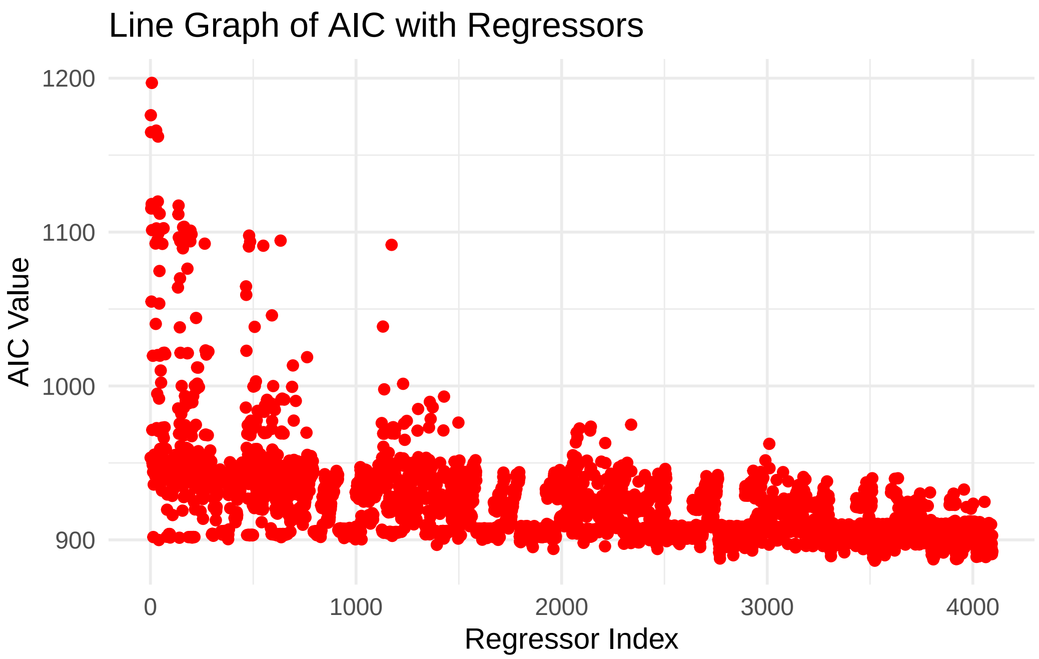
```r
generateModels <- function(response, predictors, data) {
  # Get all subsets of predictors
  attach(data)
  predictorSubsets <- getSubsets(predictors)

  # Initialize a list to store the models
  models <- list()

  # For each subset of predictors...
  for (i in 1:length(predictorSubsets)) {
    # Skip if the subset is empty
    if (length(predictorSubsets[[i]]) == 0) next

    # Generate the formula for the linear model
    formula <- as.formula(paste(response, "~", paste(predictorSubsets[[i]], collapse = "+")))

    # Fit the linear model and store it in the list
    model <- try(lm(formula, data=data), silent=TRUE)
    if (!inherits(model, "try-error")) {
      models[[length(models) + 1]] <- model
    }
  }

  return(models)
}
```

With this function we are able to get all 4095 linear models at once.

By iterating through all the death variables, we can calculate the log likelihood and AIC for all 4095 models and get the model with the highest log likelihood and the lowest AIC.

Firstly I have accounted the generate statement of more regressors are better for the model by calculating the log likelihood and AIC for all causes death variable.





The graph show that as the number of regressors increase, the log likelihood increases, and AIC decreases.

Now we get calculate the log likelihood and AIC for all possible model combinations and get the best models for each death variable.

```
   Cause Death Max LLK Index           Max LLK Min AIC Index          Min AIC
1           AC          4095 -432.229074468467          3523  886.46642803899
2           AD          4095 -307.006441841296          3799 636.243635027228
3           MN          4095 -310.914740235655          1807 641.286403470859
4         CLRD          4095 -285.486728237856          3928 593.132498822032
5           DM          4095 -277.155150497466          3864 578.413438463198
6           AH          4095 -211.595097497056          3917 446.457221523466
7          DOH          4095 -345.469410584662          3921 714.402517660974
8        EHHRD          4095 -258.463587889568          4063 541.887784976872
9           AU          4095 -335.514602444474          1404 688.218578930914
10        CLDC          4095 -256.293752604786          3800  536.98191095797
11        NNSN          4095 -281.228891862194          1021 581.165796693494
12          PD          4095  -253.53443239722          1461 525.484025972367
13          IP          4095 -286.345851612926          1535 592.388159608293
14          CD          4095 -295.984939214693           706  608.72957086489
15        ISHS          4095 -254.511508690224           791 526.297451937778
```

```
   Cause Death Max LLK Index          Max LLK Min AIC Index           Min AIC
1           AC          4095 113.625284699425          3523 -205.242290296794
2           AD          4095 69.9310559156144          3799 -117.631360486593
3           MN          4095 134.080279128775          1807 -248.703635258002
4         CLRD          4095 61.4006969772283          3928 -100.642351608138
5           DM          4095 83.0894006788725          3864  -142.07566388948
6           AH          4095 65.6060760898831          3917 -107.945125650413
7          DOH          4095 108.798329439879          3921 -194.132962388108
8        EHHRD          4095  65.184832862275          4063 -105.409056526813
9           AU          4095 45.9324280701811          1404 -74.6754820983962
10        CLDC          4095 70.9500841964467          3800 -117.505762644494
11        NNSN          4095 30.4374035852289          1021  -42.166794201353
12          PD          4095 33.3391110559168          1461 -48.2630609339061
13          IP          4095 32.3410963507594          1535 -44.9857363190789
14          CD          4095 77.6051391232049           706 -138.450585810905
15        ISHS          4095 28.6273410343593           791 -39.9802475113879
```

The first data frame is calculating log likelihood and AIC with the original data and The second data frame is calculating log likelihood and AIC with the standardized data. After doing both, the index of the best models did not change and decided to use the original data instead of the standardized data for the analysis.

The best models for each cause of death variable can be accessed by the index. By doing the calculations, we can see that all of death variables have a different combination of demographic variables for their best model. Now we are able to the analysis on the coefficients and R-squared values for these death variables. For the analysis we used the model with the minimum AIC score rather than maximum log likelihood score.

Analysis

Now that we were able to get the best possible model for each cause of death variables, we can look at the summaries of the linear models and see what is interesting about them

```r
# 1. All Causes
min.aic.lm <- all.lm[[1]][[3523]]
print(summary(min.aic.lm))
```

```

Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-868.17 -197.33  -50.27  154.58 1430.52 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -71.970463  81.912936  -0.879 0.383895    
P             0.088238   0.019404   4.547 3.58e-05 ***
A            -0.098295   0.020453  -4.806 1.50e-05 ***
B            -0.080252   0.021163  -3.792 0.000411 ***
H            -0.097149   0.020587  -4.719 2.01e-05 ***
MR           -0.215823   0.044838  -4.813 1.46e-05 ***
W            -0.085823   0.019797  -4.335 7.23e-05 ***
X0_5          0.042385   0.018387   2.305 0.025436 *  
X18_64        0.021527   0.007163   3.005 0.004173 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 461.7 on 49 degrees of freedom
Multiple R-squared:  0.9988,	Adjusted R-squared:  0.9986 
F-statistic:  4990 on 8 and 49 DF,  p-value: < 2.2e-16
```

```r
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "P"           "A"           "B"           "H"
[6] "MR"          "W"           "X0_5"        "X18_64"
```

```
# 2. "Alzheimer-Disease"
min.aic.lm <- all.lm[[2]][[3799]]
print(summary(min.aic.lm))
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-118.188  -24.004    6.546   27.502  151.075

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.788e+01  1.072e+01  -3.533 0.000919 ***
P            1.482e-02  2.823e-03   5.251 3.41e-06 ***
AI          -1.887e-02  6.458e-03  -2.921 0.005298 **
A           -1.413e-02  3.078e-03  -4.591 3.19e-05 ***
B           -1.343e-02  3.250e-03  -4.134 0.000142 ***
H           -1.342e-02  2.964e-03  -4.526 3.97e-05 ***
MR          -1.290e-02  4.344e-03  -2.969 0.004653 **
HPI         -1.683e-02  9.023e-03  -1.865 0.068292 .
W           -1.295e-02  3.025e-03  -4.282 8.84e-05 ***
X18_64      -1.591e-03  8.839e-04  -1.800 0.078093 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.03 on 48 degrees of freedom
Multiple R-squared:  0.9946,    Adjusted R-squared:  0.9936
F-statistic: 988.8 on 9 and 48 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
 [1] "(Intercept)" "P"           "AI"          "A"           "B"
 [6] "H"           "MR"          "HPI"         "W"           "X18_64"
```

```r
# 3. Malignant Neoplasms
min.aic.lm <- all.lm[[3]][[1807]]

print(summary(min.aic.lm))
```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-140.878  -27.330   -2.538   22.861  148.447 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -34.442419   9.416426  -3.658 0.000603 ***
P             0.007201   0.001796   4.011 0.000198 ***
A            -0.006032   0.001842  -3.274 0.001907 ** 
B            -0.005819   0.002201  -2.643 0.010876 *  
H            -0.005108   0.001666  -3.065 0.003472 ** 
W            -0.004336   0.001854  -2.338 0.023328 *  
X0_5         -0.011757   0.001596  -7.366 1.44e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.59 on 51 degrees of freedom
Multiple R-squared:  0.9994,	Adjusted R-squared:  0.9993 
F-statistic: 1.368e+04 on 6 and 51 DF,  p-value: < 2.2e-16
```

```r
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "P"           "A"           "B"           "H"          
[6] "W"           "X0_5"       
```

```r
# 4. Chronic lower respiratory disease
min.aic.lm <- all.lm[[4]][[3928]]

print(summary(min.aic.lm))
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-79.587 -17.077   0.648  14.593 102.965 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.630e+01  6.842e+00  -2.383 0.021181 *  
P            6.577e-03  1.575e-03   4.175 0.000125 ***
A           -6.129e-03  1.566e-03  -3.913 0.000287 ***
B           -4.756e-03  1.619e-03  -2.937 0.005081 ** 
H           -6.060e-03  1.568e-03  -3.864 0.000334 ***
MR          -1.370e-02  3.487e-03  -3.929 0.000272 ***
W           -5.318e-03  1.523e-03  -3.491 0.001043 ** 
X0_5         5.817e-03  3.052e-03   1.906 0.062628 .  
X6_17       -2.728e-03  1.219e-03  -2.238 0.029878 *  
X65_plus    -1.828e-03  7.687e-04  -2.379 0.021408 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.57 on 48 degrees of freedom
Multiple R-squared:  0.9928,	Adjusted R-squared:  0.9915 
F-statistic: 736.4 on 9 and 48 DF,  p-value: < 2.2e-16
```

```r
print(names(min.aic.lm$coefficients))
```

```
 [1] "(Intercept)" "P"           "A"           "B"           "H"          
 [6] "MR"          "W"           "X0_5"        "X6_17"       "X65_plus"   
```

```r
# 5. Diabetes Mellitus
min.aic.lm <- all.lm[[5]][[3864]]

print(summary(min.aic.lm))
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-74.200 -16.822   2.692  16.649  78.063 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.3755772  6.7853732  -0.350  0.72779    
P            0.0033965  0.0004317   7.868 3.49e-10 ***
AI          -0.0127350  0.0040341  -3.157  0.00275 ** 
A           -0.0026488  0.0003519  -7.528 1.15e-09 ***
H           -0.0033703  0.0004937  -6.826 1.36e-08 ***
MR          -0.0135296  0.0027879  -4.853 1.33e-05 ***
W           -0.0017563  0.0002686  -6.540 3.74e-08 ***
X0_5        -0.0146221  0.0026807  -5.455 1.69e-06 ***
X6_17        0.0086608  0.0010620   8.155 1.28e-10 ***
X65_plus    -0.0038363  0.0007129  -5.381 2.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.21 on 48 degrees of freedom
Multiple R-squared:  0.9965,    Adjusted R-squared:  0.9958 
F-statistic:  1510 on 9 and 48 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
 [1] "(Intercept)" "P"           "AI"          "A"           "H"          
 [6] "MR"          "W"           "X0_5"        "X6_17"       "X65_plus"   
```

```
# 6. Assult Homicide
min.aic.lm <- all.lm[[6]][[3917]]

print(summary(min.aic.lm))
```

```

Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.006  -4.748  -1.598   4.274  29.254 

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.6318311  1.8425692   2.514 0.015353 *  
P           -0.0021113  0.0005296  -3.987 0.000227 ***
A            0.0020827  0.0005388   3.865 0.000333 ***
B            0.0036190  0.0005640   6.416 5.78e-08 ***
H            0.0016903  0.0005289   3.196 0.002467 ** 
MR          -0.0040520  0.0010288  -3.939 0.000265 ***
HPI          0.0048168  0.0017196   2.801 0.007320 ** 
W            0.0022105  0.0005352   4.130 0.000144 ***
X0_5        -0.0042341  0.0006181  -6.850 1.25e-08 ***
X6_17        0.0035235  0.0003399  10.365 7.72e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.33 on 48 degrees of freedom
Multiple R-squared:  0.9937,	Adjusted R-squared:  0.9925 
F-statistic: 835.9 on 9 and 48 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
 [1] "(Intercept)" "P"           "A"           "B"           "H"          
 [6] "MR"          "HPI"         "W"           "X0_5"        "X6_17"      
```

```
# 7. Disease of hear
min.aic.lm <- all.lm[[7]][[3921]]

print(summary(min.aic.lm))
```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-244.279  -36.688    0.696   38.069  202.869 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -56.036038  18.001986  -3.113 0.003120 ** 
P             0.022762   0.005234   4.349 7.10e-05 ***
A            -0.018604   0.005286  -3.520 0.000957 ***
```

```
B             -0.016328   0.005620  -2.905 0.005536 **
H             -0.017566   0.005155  -3.408 0.001335 **
MR            -0.054502   0.010054  -5.421 1.90e-06 ***
HPI            0.035856   0.017176   2.088 0.042165 *
W             -0.015769   0.005244  -3.007 0.004192 **
X6_17         -0.013593   0.002621  -5.185 4.27e-06 ***
X65_plus      -0.005497   0.001555  -3.535 0.000915 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 104 on 48 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9983
F-statistic:  3712 on 9 and 48 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
 [1] "(Intercept)" "P"           "A"           "B"           "H"
 [6] "MR"          "HPI"         "W"           "X6_17"       "X65_plus"
```

```
# 8. Essential Hypertension
min.aic.lm <- all.lm[[8]][[4063]]

print(summary(min.aic.lm))
```

```

Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-72.082 -11.201   3.704   9.665  61.241

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.016e+01  4.416e+00  -2.302  0.02584 *
P            3.260e-03  1.198e-03   2.722  0.00908 **
A           -2.776e-03  1.231e-03  -2.255  0.02883 *
B           -2.207e-03  1.289e-03  -1.713  0.09330 .
H           -3.179e-03  1.204e-03  -2.640  0.01123 *
MR          -3.482e-03  2.338e-03  -1.489  0.14310
HPI         -1.212e-02  3.971e-03  -3.052  0.00374 **
```

```
W           -2.753e-03  1.229e-03  -2.240  0.02987 *  
X0_5        -6.554e-03  2.007e-03  -3.265  0.00204 ** 
X6_17        3.549e-03  7.875e-04   4.507 4.36e-05 ***
X65_plus    -1.549e-03  5.013e-04  -3.090  0.00336 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.35 on 47 degrees of freedom
Multiple R-squared:  0.9936,    Adjusted R-squared:  0.9922 
F-statistic: 729.8 on 10 and 47 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
 [1] "(Intercept)" "P"           "A"           "B"           "H"          
 [6] "MR"          "HPI"         "W"           "X0_5"        "X6_17"      
[11] "X65_plus"   
```

```
# 9. Accident Unintentional
min.aic.lm <- all.lm[[9]][[1404]]

print(summary(min.aic.lm))
```

```

Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-164.21  -29.18   -1.27   21.15  413.57 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.2066732 13.4379620  -0.015 0.987788    
A           -0.0011509  0.0002167  -5.311 2.30e-06 ***
H           -0.0012647  0.0003355  -3.770 0.000419 ***
MR          -0.0061525  0.0034863  -1.765 0.083472 .  
X18_64       0.0031378  0.0005427   5.782 4.25e-07 ***
X65_plus    -0.0027332  0.0006988  -3.911 0.000268 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 85.45 on 52 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9858
F-statistic: 794.9 on 5 and 52 DF,  p-value: < 2.2e-16
```

```r
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "A"           "H"           "MR"          "X18_64"
[6] "X65_plus"
```

```r
# 10. Chronic Liver
min.aic.lm <- all.lm[[10]][[3800]]

print(summary(min.aic.lm))
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-46.331 -12.774   4.029  12.548  57.985 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.3390504  4.5886283  -2.035 0.047366 *  
P            0.0049220  0.0012483   3.943 0.000261 ***
AI          -0.0064912  0.0030253  -2.146 0.036987 *  
A           -0.0045192  0.0012602  -3.586 0.000784 ***
B           -0.0035234  0.0013045  -2.701 0.009526 ** 
H           -0.0045621  0.0012329  -3.700 0.000554 ***
MR          -0.0094372  0.0021972  -4.295 8.46e-05 ***
HPI         -0.0051277  0.0038423  -1.335 0.188327    
W           -0.0040951  0.0012293  -3.331 0.001669 ** 
X65_plus    -0.0014538  0.0003478  -4.180 0.000123 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.54 on 48 degrees of freedom
Multiple R-squared:  0.9946,    Adjusted R-squared:  0.9936
F-statistic: 986.7 on 9 and 48 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
 [1] "(Intercept)" "P"           "AI"          "A"           "B"          
 [6] "H"           "MR"          "HPI"         "W"           "X65_plus"   
```

```
# 11. Nephritis
min.aic.lm <- all.lm[[11]][[1021]]

print(summary(min.aic.lm))
```

```

Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-89.059  -5.091   4.429  11.145 137.074 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.263e+00  5.344e+00  -1.359  0.17995    
P            6.318e-04  8.688e-05   7.272 1.82e-09 ***
B            7.045e-04  1.594e-04   4.418 5.07e-05 ***
MR          -8.503e-03  1.160e-03  -7.330 1.47e-09 ***
HPI          1.197e-02  4.089e-03   2.928  0.00505 ** 
X6_17       -2.248e-03  4.561e-04  -4.930 8.80e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.95 on 52 degrees of freedom
Multiple R-squared:  0.9774,	Adjusted R-squared:  0.9752 
F-statistic: 449.4 on 5 and 52 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "P"           "B"           "MR"          "HPI"        
[6] "X6_17"      
```

```
# 12. Parkinsons
min.aic.lm <- all.lm[[12]][[1461]]

print(summary(min.aic.lm))
```

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

Residuals:
    Min      1Q  Median      3Q     Max 
-46.854  -6.379   5.632  14.687  42.645 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.654e+01  3.325e+00  -4.975 7.51e-06 ***
B           -9.182e-04  1.592e-04  -5.766 4.50e-07 ***
H            4.077e-04  8.123e-05   5.019 6.45e-06 ***
MR           7.170e-03  9.616e-04   7.456 9.27e-10 ***
HPI         -9.227e-03  2.166e-03  -4.261 8.58e-05 ***
X0_5        -2.757e-03  7.796e-04  -3.537 0.000862 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.01 on 52 degrees of freedom
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9777 
F-statistic: 500.3 on 5 and 52 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "B"           "H"           "MR"          "HPI"        
[6] "X0_5"       
```

```
# 13. Influenza
min.aic.lm <- all.lm[[13]][[1535]]

print(summary(min.aic.lm))
```

```
Call:
```

```
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-96.184 -13.387   9.297  13.046 139.770 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.145e+01  5.905e+00  -1.939  0.05787 .  
H            3.216e-04  1.088e-04   2.956  0.00468 ** 
MR          -3.505e-03  2.040e-03  -1.718  0.09169 .  
HPI          1.092e-02  4.298e-03   2.540  0.01409 *  
X0_5        -6.376e-03  9.641e-04  -6.613 2.05e-08 ***
X18_64       8.104e-04  1.325e-04   6.115 1.27e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.4 on 52 degrees of freedom
Multiple R-squared:  0.9785,    Adjusted R-squared:  0.9764 
F-statistic: 472.2 on 5 and 52 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "H"           "MR"          "HPI"         "X0_5"       
[6] "X18_64"     
```

```
# 14. Cerebrovascular
min.aic.lm <- all.lm[[14]][[706]]

print(summary(min.aic.lm))
```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-101.249  -25.501    2.541   17.449  111.983 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
```

```
(Intercept) -2.406e+01  6.991e+00  -3.441 0.001139 **
B            6.655e-04  2.553e-04   2.607 0.011839 *
HPI          1.072e-02  2.729e-03   3.929 0.000249 ***
W            7.494e-04  9.060e-05   8.271 4.14e-11 ***
X18_64       2.458e-04  6.164e-05   3.987 0.000206 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.4 on 53 degrees of freedom
Multiple R-squared:  0.9955,	Adjusted R-squared:  0.9952
F-statistic:  2963 on 4 and 53 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "B"           "HPI"         "W"           "X18_64"
```

```
# 15. Intentional
min.aic.lm <- all.lm[[15]][[791]]

print(summary(min.aic.lm))
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-52.828 -14.334   8.561  14.877  38.791

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.808e+00  3.613e+00  -2.438  0.01817 *
W            2.104e-04  3.707e-05   5.675 5.92e-07 ***
X0_5        -2.768e-03  8.191e-04  -3.379  0.00137 **
X18_64       6.989e-04  1.437e-04   4.864 1.07e-05 ***
X65_plus    -1.202e-03  2.506e-04  -4.796 1.36e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.32 on 53 degrees of freedom
Multiple R-squared:  0.9757,	Adjusted R-squared:  0.9739
F-statistic: 531.7 on 4 and 53 DF,  p-value: < 2.2e-16
```

```
print(names(min.aic.lm$coefficients))
```

```
[1] "(Intercept)" "W"           "X0_5"        "X18_64"      "X65_plus"
```

There are 15 different cause of death variables that were considered and we are not able to analyze all thoroughly. From the model selected based on the AIC values, I found that the Malignant Neoplasms and Intentional Self Harm was the most interesting.

Malignant Neoplasms

The reason for the Malignant Neoplasms variable was that all of the variables were race variables besides one age variable. The best model variables are:

- LM (MN ~ P + A + B + H + W + X0_5)
- Variables : total population, Asian, Black, Hispanic, White, ages 0-5

Malignant Neoplasm is another term for a cancerous tumor. About 56.1 percent of cases are diagnosed in those older than 65 *1.

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

Residuals:
     Min       1Q   Median       3Q      Max 
-140.878  -27.330   -2.538   22.861  148.447 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -34.442419   9.416426  -3.658 0.000603 ***
P             0.007201   0.001796   4.011 0.000198 ***
A            -0.006032   0.001842  -3.274 0.001907 ** 
B            -0.005819   0.002201  -2.643 0.010876 *  
H            -0.005108   0.001666  -3.065 0.003472 ** 
W            -0.004336   0.001854  -2.338 0.023328 *  
X0_5         -0.011757   0.001596  -7.366 1.44e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.59 on 51 degrees of freedom
Multiple R-squared:  0.9994,	Adjusted R-squared:  0.9993 
F-statistic: 1.368e+04 on 6 and 51 DF,  p-value: < 2.2e-16
```

```
[1] "(Intercept)" "P"           "A"           "B"           "H"          
[6] "W"           "X0_5"       
```

Based on the best model for the death variable Malignant Neoplasms, it was surprising that the age 65+ variable was not a regressor for the best model for Malignant Neoplasms. I wanted to further analyze by replacing the age 0-5 variable with age 0-5 and see what the the outcome would have been.

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

Residuals:
     Min       1Q   Median       3Q      Max 
-140.878  -27.330   -2.538   22.861  148.447 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -34.442419   9.416426  -3.658 0.000603 ***
P             0.007201   0.001796   4.011 0.000198 ***
A            -0.006032   0.001842  -3.274 0.001907 ** 
B            -0.005819   0.002201  -2.643 0.010876 *  
H            -0.005108   0.001666  -3.065 0.003472 ** 
W            -0.004336   0.001854  -2.338 0.023328 *  
X0_5         -0.011757   0.001596  -7.366 1.44e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.59 on 51 degrees of freedom
Multiple R-squared:  0.9994,    Adjusted R-squared:  0.9993 
F-statistic: 1.368e+04 on 6 and 51 DF,  p-value: < 2.2e-16

  (Intercept)             P             A             B             H 
-34.442419258   0.007201116  -0.006031838  -0.005819371  -0.005107640 
            W          X0_5 
 -0.004336041  -0.011756804 
```

```
Call:
lm(formula = data.set$MN ~ data.set$P + data.set$A + data.set$B + 
    data.set$H + data.set$W + data.set$X65_plus)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-173.863  -32.669    1.798   35.750  125.004

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -4.321e+01  1.080e+01  -4.001 0.000204 ***
data.set$P         4.304e-03  1.839e-03   2.340 0.023258 *
data.set$A        -4.386e-03  2.013e-03  -2.179 0.033992 *
data.set$B        -3.692e-03  2.390e-03  -1.545 0.128546
data.set$H        -3.581e-03  1.816e-03  -1.972 0.054069 .
data.set$W        -2.943e-03  2.050e-03  -1.436 0.157240
data.set$X65_plus  3.610e-03  6.357e-04   5.679 6.51e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63.63 on 51 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9991
F-statistic: 1.082e+04 on 6 and 51 DF,  p-value: < 2.2e-16

      (Intercept)        data.set$P        data.set$A        data.set$B
    -43.207032823       0.004303539      -0.004386475      -0.003692356
       data.set$H        data.set$W data.set$X65_plus
     -0.003581306      -0.002942786       0.003609701

The AIC with X0_5:  641.2864

The AIC with X65_plus:  654.8948

R Squared

The R Squared with X0_5:  0.9993791

The R Squared with X65_plus:  0.9992149

data.set$X65_plus
      0.003609701

      X0_5
-0.0117568
```

Based on the summary of both linear model of with age 0-5 and age 65+, we can still say that the model with age 0-5 is a better model as the AIC is lower and the R-squared value is higher than the model with age 65+.

Results were:

- The AIC with X0_5: 641.2864 *** Better Model**
- The AIC with X65_plus: 654.8948
- The R Squared with X0_5: 0.9993791 *** Better Fit**
- The R Squared with X65_plus: 0.9992149

By doing a bootstrap, the results of the coefficient of the variables ages 0-5 and ages 65+ were :

- The 99th confidence interval for coefficient for age 0-5 is: -0.02030863 - 0.008203702
- The 99th confidence interval for coefficient for age 65+ is: -0.002102959 - 0.006924535

Since the lower the AIC the better, we still see that the model with the age 0-5 variable is a better model as it has a lower AIC than the other model. Even though the R-squared values of both variables are very high, the model with age 0-5 variable is higher. From these two statistics, we can state that the model with ages 0-5 variable is a better model in terms of fit and complexity and explains the variance in the dependent variables a bit better than the model with ages 0-5 variable. The confidence interval for coefficients for age 0-5 and age 65+ both include zero, which indicates that despite the p-value indicating the high significance, both variables may have a chance of not being statistically significant.

Intentional Self Harm Suicide

The reason for the analysis of Intentional Self Harm Suicide variable was that all of the variables were age variables besides one race variable. The best model variables are:

- LM (ISHS ~ W + X0_5 + X18_64 + X65_plus)
- Variables : White, ages 0-5, ages 18-64, ages 65+

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

Residuals:
    Min      1Q  Median      3Q     Max
-52.828 -14.334   8.561  14.877  38.791
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.808e+00  3.613e+00  -2.438  0.01817 *
W            2.104e-04  3.707e-05   5.675 5.92e-07 ***
X0_5        -2.768e-03  8.191e-04  -3.379  0.00137 **
X18_64       6.989e-04  1.437e-04   4.864 1.07e-05 ***
X65_plus    -1.202e-03  2.506e-04  -4.796 1.36e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.32 on 53 degrees of freedom
Multiple R-squared:  0.9757,    Adjusted R-squared:  0.9739
F-statistic: 531.7 on 4 and 53 DF,  p-value: < 2.2e-16
```

```
[1] "(Intercept)" "W"           "X0_5"        "X18_64"      "X65_plus"
```

It was surprising that all of the dependent variables were age variables besides one. We can further analyze the white demographic variable.

Based on the summary the model, we can still say that the white variable is statistically significant as the p-value of W is 5.92e-07 which is smaller than the significance level of 0.05.

```
The 99th confidence interval for coefficient for age 0-5 is:  -0.01979355 0.005743655
```

```
The 99th confidence interval for coefficient for age 65+ is:  -0.001604217 0.006652516
```

By doing a bootstrap, the results of the coefficient of the variables ages white were :

- The 99th confidence interval for coefficient for white is: -0.000169536 - 0.000362979

The confidence interval for coefficients for white include zero, which indicates that despite the p-value indicating the high significance, white variables may have a chance of not being statistically significant.

Conclusions

In this project, we conducted an extensive analysis of the relationship between various demographic factors and causes of death across California's 58 counties. Our primary objective was to understand how different demographic groups impact mortality patterns for different health conditions.

We sourced our data from the California Department of Public Health and the California State Association of Counties, combining death records from 2019-2021 with demographic data from 2023. The analysis involved significant data cleaning and integration, culminating in a data set that allowed us to perform detailed statistical analyses using linear regression models.

Key Findings:

1. **Model Selection**: By using log likelihood (LLK) and Akaike Information Criterion (AIC) as metrics, we identified the best-fit models for each cause of death. The use of these metrics ensured that our models balanced fit and complexity effectively.

2. **Exploratory Data Analysis (EDA)**: Heat maps highlighted the population distribution and mortality rates across counties, providing a visual representation of demographic and health disparities.

3. **Significant Demographic Factors**:

   - For **Malignant Neoplasms (cancerous tumors)**, the best model included the total population, Asian, Black, Hispanic, White demographics, and the 0-5 age group. Interestingly, the expected 65+ age group was not part of the best model. Further analysis confirmed that the 0-5 age group model had a lower AIC and higher R-squared value, indicating a better fit.

   - For **Intentional Self-Harm (Suicide)**, the best model included the White demographic and three age groups (0-5, 18-64, and 65+). The White demographic was found to be statistically significant, though the confidence interval suggested potential non-significance.

4. **Model Analysis**

   - The models for **Malignant Neoplasms** and **Intentional Self-Harm** showed that certain demographic factors are more influential than others. For instance, age demographics played a crucial role in suicide rates, while race demographics were more significant for cancer mortality.

5. **Bootstrapping**: This technique was employed to assess the stability and reliability of our model coefficients. For both Malignant Neoplasms and Intentional Self-Harm, the confidence intervals for key demographic coefficients included zero, suggesting potential non-significance despite low p-values in the regression models.

6. **Limitations**: The analysis faced challenges such as significant population disparities among counties and potential biases in demographic data. Additionally, high correlation among variables made it difficult to pinpoint the exact influence of individual demographics on mortality rates.

Overall, this study highlights the complex interplay between demographics and mortality causes in California. The insights gained can inform public health strategies and targeted interventions to address specific health risks within different demographic groups. Future research could expand on these findings by incorporating additional variables and exploring non-linear models to capture more nuanced relationships.

Appendix

Initial cleaning in Python

```
''' Initial data cleaning in jupyter notebook in python'''

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math

''' Final Death by Month by Occurrence '''
final_death_occurrence = pd.read_csv('death-profiles-by-county-bzbdqj/2019-2021-final-deaths-

## combining count and annotation code to get the final count
count = final_death_occurrence['Count'].to_list()
annotation = final_death_occurrence['Annotation_Code'].to_list()

final_count = []
for i in range(len(count)):
    if math.isnan(count[i]):
        final_count.append(annotation[i])
    else:
        final_count.append(count[i])

# print(any(math.isnan(i) for i in final_count))
# print(final_count[1:6])

# Drop unnecessary columns
modified_final_death_occurrence = final_death_occurrence.drop(columns=['Geography_Type', 'IC
modified_final_death_occurrence["Count"] = final_count

# Get all the unique counties
modified_final_death_occurrence['County'].unique()

# Get all the unique causes of death
```

```
modified_final_death_occurrence['Cause_Desc'].unique()

# Example : Total deaths in Alameda County in 2021 all causes of death
grouped = modified_final_death_occurrence.groupby(['Year', 'County', 'Strata', 'Cause_Desc']]
group_2019_Alameda_TotalPop = grouped.get_group((2021, 'Alameda', 'Total Population', 'All ca
print(group_2019_Alameda_TotalPop['Count'].sum())
group_2019_Alameda_TotalPop

''' Demographic Data from 2023'''

# Read demographic data
demo_data_2023 = pd.read_excel('demographic.xlsx', sheet_name=None)

# Get the "people" excel sheet
people_data = demo_data_2023["People"]

# list of headers
people_header = people_data.columns.tolist()

# Drop unnecessary columns
people_data = people_data.drop(columns=[ 'Population in Group Quarters (January 2023)',
 'Population: Unincorporated (January 2023)',
 'Population: Unincorporated, Group Quarters (January 2023)',
 'Eligible Voters (February 2023)',
 'Registered Voters: No Party Preference (February 2023)',
 'Registered Voters: Democrat (February 2023)',
 'Registered Voters: Republican (February 2023)',
 'Registered Voters: Other Party (February 2023)',
 'Labor Force (9-2023)',
 'Labor Force: Employed (9-2023)',
 'Labor Force: Unemployed (9-2023)',
 'Labor Force: Unemployment Rate (9-2023)',
 'Poverty: All Ages (2021)',
 'Poverty: All Ages Percent (2021)',
 'Poverty: Under 18 (2021)',
 'Poverty: Under 18 Percent (2021)',
 'Median Household Income (2021)'])

# unique values for strata name column
unique_values = modified_final_death_occurrence['Strata_Name'].unique()

# print our first 5 rows
```

```
people_data[0:5]

# List of all the causes of death and counties
Cause_of_death = ['All causes (total)', "Alzheimer's disease", 'Malignant neoplasms',
       'Chronic lower respiratory diseases', 'Diabetes mellitus',
       'Assault (homicide)', 'Diseases of heart',
       'Essential hypertension and hypertensive renal disease',
       'Accidents (unintentional injuries)',
       'Chronic liver disease and cirrhosis',
       'Nephritis, nephrotic syndrome and nephrosis',
       "Parkinson's disease", 'Influenza and pneumonia',
       'Cerebrovascular diseases', 'Intentional self-harm (suicide)']

Counties = ['Alameda', 'Alpine', 'Amador', 'Butte', 'Calaveras', 'Colusa',
       'Contra Costa', 'Del Norte', 'El Dorado', 'Fresno', 'Glenn',
       'Humboldt', 'Imperial', 'Inyo', 'Kern', 'Kings', 'Lake', 'Lassen',
       'Los Angeles', 'Madera', 'Marin', 'Mariposa', 'Mendocino',
       'Merced', 'Modoc', 'Mono', 'Monterey', 'Napa', 'Nevada', 'Orange',
       'Placer', 'Plumas', 'Riverside', 'Sacramento', 'San Benito',
       'San Bernardino', 'San Diego', 'San Francisco', 'San Joaquin',
       'San Luis Obispo', 'San Mateo', 'Santa Barbara', 'Santa Clara',
       'Santa Cruz', 'Shasta', 'Sierra', 'Siskiyou', 'Solano', 'Sonoma',
       'Stanislaus', 'Sutter', 'Tehama', 'Trinity', 'Tulare', 'Tuolumne',
       'Ventura', 'Yolo', 'Yuba']

print(f"num of cause of death : {len(Cause_of_death)}")
print(f"num counties : {len(Counties)}")
print(len(people_data) == len(Counties))



# loop through the different groups in death data and gather all the data for that the county

grouped = modified_final_death_occurrence.groupby(['Year', 'County', 'Strata', 'Cause_Desc']]

i = 0
while i < len(Cause_of_death):
    column_name = "2021 " + Cause_of_death[i]
    county_death = []

    for county in Counties:
        county_data = grouped.get_group((2021, county, 'Total Population', Cause_of_death[i]]
        county_death += [county_data['Count'].sum()]

    # add column to people data df
    people_data[column_name] = county_death

    i += 1

#modified demographic + deaths data
people_data.to_csv("Data/Final Data/demographic_death.csv", index=False)

# Print column names to check final demographic/death data
people_data.columns.tolist()
```

Setup Data for R Analysis

```
# Load the data for demographic and death data
data.set <- read.csv("Data/demographic_death.csv")

# attaching variable names to columns
column.names <- c("County", "Population",
                    "American-Indian", "Asian", "Black", "Hispanic", "Multi-Racial", "Pacifi
                    "0-5", "6-17", "18-64", "65+",
                    "All-Causes", "Alzheimer-Disease", "Malignant-Neoplasms",
                    "Chronic-Lower-Respiratory-Disease", "Diabetes-Mellitus", "Assult-Homoci
                    "Disease-of-Heart",
                    "Essential-Hypertension-Hypertensive-Renal-Disease", "Accident-Unintenti
                    "Chronic-Liver-Disease-Cirrhosis", "Nephritis-Nephrotic-Syndrome-Nephrosi
                    "Parkinsons-Disease", "Influenza-Pneumonia",
                    "Cerebrovascular-Disease", "Intentional-Self-Harm-Suicide")

# Attaching variable names to columns
names(data.set) <- c("C", "P", "AI", "A", "B", "H", "MR", "HPI", "W",
                     "X0_5", "X6_17", "X18_64", "X65_plus",
                     "AC", "AD", "MN", "CLRD", "DM", "AH", "DOH", "EHHRD", "AU", "CLDC",
                     "NNSN", "PD", "IP", "CD", "ISHS")

# Demographic variables
dem <- c("P", "AI", "A", "B", "H", "MR", "HPI", "W",
         "X0_5", "X6_17", "X18_64", "X65_plus")
```

```r
# Cause of death variables
dea <- c("AC", "AD", "MN", "CLRD", "DM", "AH", "DOH", "EHHRD", "AU", "CLDC",
         "NNSN", "PD", "IP", "CD", "ISHS")


# Standarizing the data
std.data.set <- data.set
for (i in 2:ncol(data.set)) {
  std.data.set[, i] <- (data.set[, i] - mean(data.set[,i], na.rm=TRUE))
  std.data.set[, i] <- std.data.set[,i]/sd(data.set[,i], na.rm=TRUE)
}
attach(std.data.set)

#### data.set -> original data
#### std.data.set -> standardized data
```

Exploratory data analysis:

```r
### Heat Maps

# total population based on the counties
library(ggplot2)
library(sf)
library(tigris)
library(dplyr)
library(viridis)

options(tigris_class = "sf")

# Get ca counties information
ca_counties <- counties(state = "CA", cb = TRUE, class = "sf")

# order the counties in alphabetical along
ca_counties <- arrange(ca_counties, NAME)

# add population data to the ca_counties data frame
ca_counties <- ca_counties %>%
  mutate(population = data.set$P)

# Create the heat map
ggplot(data = ca_counties) +
  geom_sf(aes(fill = population), color = "white", size = 0.1) +
  scale_fill_viridis_c(option = "magma", name = "Population") +
  labs(title = "Population Heat Map of California Counties") +
  theme_minimal()

# total Death by Alzheimer Disease based on the counties
library(ggplot2)
library(sf)
library(tigris)
library(dplyr)
library(viridis)

options(tigris_class = "sf")

# Get ca counties information
ca_counties <- counties(state = "CA", cb = TRUE, class = "sf")

# order the counties in alphabetical along
ca_counties <- arrange(ca_counties, NAME)

# add population data to the ca_counties data frame
ca_counties <- ca_counties %>%
  mutate(population = data.set$AD)

# Create the heat map
ggplot(data = ca_counties) +
  geom_sf(aes(fill = population), color = "white", size = 0.1) +
  scale_fill_viridis_c(option = "magma", name = "Population") +
  labs(title = "Death by Alzheimer's Disease Heat Map of California Counties") +
  theme_minimal()

# total population of White Californians based on the counties
library(ggplot2)
library(sf)
library(tigris)
library(dplyr)
library(viridis)

options(tigris_class = "sf")

# Get ca counties information
ca_counties <- counties(state = "CA", cb = TRUE, class = "sf")
```

```r
# order the counties in alphabetical along
ca_counties <- arrange(ca_counties, NAME)

# add population data to the ca_counties data frame
ca_counties <- ca_counties %>%
  mutate(population = data.set$W)

# Create the heat map
ggplot(data = ca_counties) +
  geom_sf(aes(fill = population), color = "white", size = 0.1) +
  scale_fill_viridis_c(option = "magma", name = "Population") +
  labs(title = "White Population Heat Map of California Counties") +
  theme_minimal()
```

Model Selection and Linear Model

```r
# Varius function for analysis

# function to calculate log likelihood for a linear model
calc.llk.lm <- function(lm.in){
  eps <- lm.in$residuals
  N <- length(eps)
  var <- 1/(N) *sum(eps^2, na.rm=TRUE)
  llk <- sum(dnorm(eps, 0, sqrt(var), log=TRUE), na.rm=TRUE)
  return (llk)
}

# Function to calculate the AIC
calc.aic <- function(model){
  # Getting number of parameters
  K <-length(coef(model)) + 1


  # Getting the log likelihood
  llk <- calc.llk.lm(model)

  aic <- 2 * K - 2 * llk
  return (aic)
}

## Optional : Function to calculate the number of parameters in a linear model
```

```r
calc.num.parameters <- function(lm.in){
  K <-length(coef(lm.in)) + 1
  return (K)
}


# Function to generate all possible subsets of a vector
getSubsets <- function(vec) {
  n <- length(vec)
  subsets <- list()
  for (i in 0:n) {
    subsets <- c(subsets, combn(vec, i, simplify = FALSE))
  }
  return(subsets)
}

# Function to generate all possible linear models
generateModels <- function(response, predictors, data) {
  # Get all subsets of predictors
  attach(data)
  predictorSubsets <- getSubsets(predictors)

  # Initialize a list to store the models
  models <- list()

  # For each subset of predictors...
  for (i in 1:length(predictorSubsets)) {
    # Skip if the subset is empty
    if (length(predictorSubsets[[i]]) == 0) next

    # Generate the formula for the linear model
    formula <- as.formula(paste(response, "~", paste(predictorSubsets[[i]], collapse = "+")))

    # Fit the linear model and store it in the list
    model <- try(lm(formula, data=data), silent=TRUE)
    if (!inherits(model, "try-error")) {
      models[[length(models) + 1]] <- model
    }
  }

  return(models)
}
```

```r
### Example of LLK and AIC plot for all causes death variable
# All Causes
library(ggplot2)
all.causes <- dea[1]
regressors <- dem
models <- generateModels(all.causes, regressors, data.set)

llk <- sapply(models, calc.llk.lm)
aic <- sapply(models, calc.aic)

# Create a data frame
df <- data.frame(
  Regressor = seq_along(aic),    # Create an index for the x-axis
  LLK = llk,                         # The LLK values for the y-axis
  AIC = aic                         # The AIC values for the y-axis
)



# Plot the line graph for LLK
ggplot(df, aes(x = Regressor, y = LLK)) +
  geom_point(color = "red") +  # Adding points to the line for better visibility
  labs(title = "Line Graph of LLK with Regressors",
       x = "Regressor Index",
       y = "LLK Value") +
  theme_minimal()

# Plot the line graph for AIC
ggplot(df, aes(x = Regressor, y = AIC)) +
  geom_point(color = "red") +  # Adding points to the line for better visibility
  labs(title = "Line Graph of AIC with Regressors",
       x = "Regressor Index",
       y = "AIC Value") +
  theme_minimal()

# Get LLK and AIC for linear models
lm.matrix <- matrix(nrow = length(dea), ncol = 5)
std.matrix <- matrix(nrow = length(dea), ncol = 5)

# store all lm
all.lm <- list()
```

```r
# iterating through all death variables
for (i in 1:length(dea)) {
  dep.var <- dea[i]
  regressors <- dem

  # Generate all possible models for data + standardized data
  models <- generateModels(dep.var, regressors, data.set)
  std.models <- generateModels(dep.var, regressors, std.data.set)

  # save models for later use
  all.lm[[i]] <- models

  # Calculate LLK for each model
  llk <- sapply(models, calc.llk.lm)
  std.llk <- sapply(std.models, calc.llk.lm)

  # Calculate AIC for each model
  aic <- sapply(models, calc.aic)
  std.aic <- sapply(std.models, calc.aic)

  # Store the results in the matrix
  maxLLKIndex <- which.max(llk)
  minAICIndex <- which.min(aic)
  std.maxLLKIndex <- which.max(std.llk)
  std.minAICIndex <- which.min(std.aic)

  # store them in model
  lm.matrix[i, 1] <- dep.var
  lm.matrix[i, 2] <- maxLLKIndex
  lm.matrix[i, 3] <- max(llk)
  lm.matrix[i, 4] <- minAICIndex
  lm.matrix[i, 5] <- min(aic)

  std.matrix[i, 1] <- dep.var
  std.matrix[i, 2] <- std.maxLLKIndex
  std.matrix[i, 3] <- max(std.llk)
  std.matrix[i, 4] <- std.minAICIndex
  std.matrix[i, 5] <- min(std.aic)
  print("-\n")
}
```

```
lm.df <- as.data.frame(lm.matrix)
colnames(lm.df) <- c("Cause Death", "Max LLK Index", "Max LLK", "Min AIC Index", "Min AIC")

std.lm.df <- as.data.frame(std.matrix)
colnames(std.lm.df) <- c("Cause Death", "Max LLK Index", "Max LLK", "Min AIC Index", "Min AIC

print(lm.df)
print(std.lm.df)
```

Best Models for each Death Variable

```
# 1. All Causes
min.aic.lm <- all.lm[[1]][[3523]]
print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 2. "Alzheimer-Disease"
min.aic.lm <- all.lm[[2]][[3799]]
print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 3. Malignant Neoplasms
min.aic.lm <- all.lm[[3]][[1807]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 4. Chronic lower respiratory disease
min.aic.lm <- all.lm[[4]][[3928]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 5. Diabetes Mellitus
min.aic.lm <- all.lm[[5]][[3864]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 6. Assult Homicide
min.aic.lm <- all.lm[[6]][[3917]]
```

```r
print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))


# 7. Disease of hear
min.aic.lm <- all.lm[[7]][[3921]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 8. Essential Hypertension
min.aic.lm <- all.lm[[8]][[4063]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 9. Accident Unintentional
min.aic.lm <- all.lm[[9]][[1404]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 10. Chronic Liver
min.aic.lm <- all.lm[[10]][[3800]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 11. Nephritis
min.aic.lm <- all.lm[[11]][[1021]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 12. Parkinsons
min.aic.lm <- all.lm[[12]][[1461]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 13. Influenza
min.aic.lm <- all.lm[[13]][[1535]]
```

```
print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 14. Cerebrovascular
min.aic.lm <- all.lm[[14]][[706]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))

# 15. Intentional
min.aic.lm <- all.lm[[15]][[791]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))
```

Analysis and Bootstrap of Malignant Neoplasms

```
# 3. Malignant Neoplasms
min.aic.lm <- all.lm[[3]][[1807]]
summary(min.aic.lm)
print(min.aic.lm$coefficients)


# 65 + model
# Assuming your data frame is named data.set and the variables are properly named
lm.65.plus <- lm(data.set$MN ~ data.set$P + data.set$A + data.set$B + data.set$H + data.set$

# Display the summary of the linear model
summary(lm.65.plus)
print(lm.65.plus$coefficients)

# comparing aic
min.aic.lm.aic <- calc.aic(min.aic.lm)
lm.65.plus.aic <- calc.aic(lm.65.plus)
cat("The AIC with X0_5: ", min.aic.lm.aic, "\n")
cat("The AIC with X65_plus: ", lm.65.plus.aic, "\n")

min.aic.lm.r.sq <- summary(min.aic.lm)$r.squared
lm.65.plus.r.sq <- summary(lm.65.plus)$r.squared
cat("R Squared \n")
cat("The R Squared with X0_5: ", min.aic.lm.r.sq, "\n")
cat("The R Squared with X65_plus: ", lm.65.plus.r.sq, "\n")

# Boostrapping
lm.65.plus$coefficients[7]
min.aic.lm$coefficients[7]

X.0.5.coef <- rep(NA, M)
X.65.plus.coef <- rep(NA, M)

for (j in 1:M){

  bs.pos <- sample(1:N, N, replace=TRUE)
  df.bs <- data.set[bs.pos, ]

  ## Recalculate
  lm.0.5.temp <- lm(df.bs$MN ~ df.bs$P + df.bs$A + df.bs$B + df.bs$H + df.bs$W + df.bs$X0_5)
  lm.65.plus.temp <- lm(df.bs$MN ~ df.bs$P + df.bs$A + df.bs$B + df.bs$H + df.bs$W + df.bs$X6

  X.0.5.coef[j] <- lm.0.5.temp$coefficients[7]
  X.65.plus.coef[j] <- lm.65.plus.temp$coefficients[7]

}

c.i.99.0.5 <- quantile(X.0.5.coef, c(0.005, 0.995))

c.i.99.65.plus <-quantile(X.65.plus.coef, c(0.005, 0.995))

cat("The 99th confidence interval for coefficient for age 0-5 is: ", c.i.99.0.5, "\n")
cat("The 99th confidence interval for coefficient for age 65+ is: ", c.i.99.65.plus, "\n")
```

Analysis and Bootstrap of Intentional Suicide

```
# 15. Intentional
min.aic.lm <- all.lm[[15]][[791]]

print(summary(min.aic.lm))
print(names(min.aic.lm$coefficients))
print(coef(min.aic.lm))
print(min.aic.lm$coefficients[2])

# Bootstrapping
```