To find the best suited multiple linear regression model for the expected cancer related death rate per county using the incidence rate of cancer and the available socio-economic data for the county, we can follow the following steps in R:

Step 1: Load the dataset

```
#Uploaded the file to the Github inorder to be pulled/viewed from any
location
cancer_data <-
read.csv("https://raw.githubusercontent.com/josephagoi/cancer-
dataset/main/cancer.csv")

#Printing the header to check the uploaded datset
head(cancer_data)
```

```
  County                         Population deathRate incidenceRate
medIncome
1 Iosco County, Michigan        25345      193.4     406.2
37122
2 Mineral County, Montana        4251      188.8     538.8
36449
3 Lake County, Oregon            7829      139.2     397.2
40328
4 Pittsylvania County, Virginia 62194      176.8     399.0
44207
5 Hall County, Texas             3138      223.2     415.8
33324
6 Lane County, Kansas            1670      148.6     371.5
53739
  povertyPercent MedianAge MedianAgeMale MedianAgeFemale
AvgHouseholdSize …
1 19.0           624.0     51.6          52.5            2.20
…
2 17.3           619.2     52.3          50.7            2.67
…
3 19.3           579.6     48.2          48.5            2.08
…
4 14.6           546.0     44.2          46.7            2.36
…
5 24.5           536.4     42.8          45.2            2.71
…
6 10.9           535.2     44.7          44.4            2.07
…
  PctBachDeg25_Over PctUnemployed16_Over PctPrivateCoverage
PctEmpPrivCoverage
1 10.0              12.1                 61.4               29.4

2  9.8              11.8                 48.3               20.2

3 12.1              10.6                 54.7               29.9
```

```
4  9.4               7.7                 66.2               42.6

5  8.9              11.5                 42.6               20.4

6 15.5               2.4                 78.2               48.6

  PctPublicCoverage PctPublicCoverageAlone PctWhite PctBlack
PctAsian
1 53.2              23.4                   96.14582  0.62595961
0.5196646
2 53.8              27.1                   98.44266  0.07078811
0.1651723
3 48.3              25.6                   90.28309  0.61208875
0.8161183
4 36.2              18.6                   74.84473 21.41765137
0.3423894
5 42.7              26.9                   81.67343  7.46175461
0.3122073
6 29.2              11.2                   98.32736  0.35842294
0.0000000
  PctOtherRace
1 0.1417267
2 0.1415762
3 2.3845958
4 1.5765837
5 6.2441461
6 0.0000000
```

Step 2: Clean and preprocess the data

```
# Remove unnecessary columns
cancer_data_clean <- cancer_data[, c(1:21)]

# Remove rows with missing values
cancer_data_clean <- na.omit(cancer_data_clean)

# Check the data structure and summary
str(cancer_data_clean)
summary(cancer_data_clean)
```

```
'data.frame':   3047 obs. of  21 variables:
 $ County                : chr  "Iosco County, Michigan" "Mineral
County, Montana" "Lake County, Oregon" "Pittsylvania County, Virginia"
...
 $ Population            : int  25345 4251 7829 62194 3138 1670 93246
3910 126517 127253 ...
 $ deathRate             : num  193 189 139 177 223 ...
 $ incidenceRate         : num  406 539 397 399 416 ...
 $ medIncome             : int  37122 36449 40328 44207 33324 53739
40429 37581 70705 47175 ...
```

```
 $ povertyPercent        : num  19 17.3 19.3 14.6 24.5 10.9 15.9 19.4
10.4 14.7 ...
 $ MedianAge             : num  624 619 580 546 536 ...
 $ MedianAgeMale         : num  51.6 52.3 48.2 44.2 42.8 44.7 41.6
42.4 42 41 ...
 $ MedianAgeFemale       : num  52.5 50.7 48.5 46.7 45.2 44.4 46.2
45.4 44.5 43.9 ...
 $ AvgHouseholdSize      : num  2.2 2.67 2.08 2.36 2.71 2.07 2.27 2.34
2.42 2.51 ...
 $ PctMarriedHouseholds  : num  48.1 46.8 47.6 51.6 51.5 ...
 $ PctNoHS18_24          : num  25.2 17 7.7 14.7 27.4 25.2 22 26.9 6.2
15.4 ...
 $ PctHS18_24            : num  32.4 59.8 54 40.7 41.8 31.1 40.2 27.6
28.5 40.6 ...
 $ PctBachDeg18_24       : num  2.2 13 4.5 6.3 0 3 7.9 13.1 12.9
5.7 ...
 $ PctHS25_Over          : num  40 41.8 33.4 35.3 27.9 29.7 50.3 41.9
23.3 34.9 ...
 $ PctBachDeg25_Over     : num  10 9.8 12.1 9.4 8.9 15.5 9.4 11 25.8
15 ...
 $ PctUnemployed16_Over  : num  12.1 11.8 10.6 7.7 11.5 2.4 7.2 7.4
6.4 7.8 ...
 $ PctPrivateCoverage    : num  61.4 48.3 54.7 66.2 42.6 78.2 71.8 56
81.7 68.7 ...
 $ PctEmpPrivCoverage    : num  29.4 20.2 29.9 42.6 20.4 48.6 46.5
28.4 57.3 45.4 ...
 $ PctPublicCoverage     : num  53.2 53.8 48.3 36.2 42.7 29.2 37 37.4
27.3 35.1 ...
 $ PctPublicCoverageAlone: num  23.4 27.1 25.6 18.6 26.9 11.2 16 18.6
11.2 17.5 ...

    County            Population         deathRate     incidenceRate

 Length:3047        Min.   :     827   Min.   : 59.7   Min.   : 201.3

 Class :character   1st Qu.:   11684   1st Qu.:161.2   1st Qu.: 413.1

 Mode  :character   Median :   26643   Median :178.1   Median : 449.5

                    Mean   :  102637   Mean   :178.7   Mean   : 445.7

                    3rd Qu.:   68671   3rd Qu.:195.2   3rd Qu.: 482.1

                    Max.   :10170292   Max.   :362.8   Max.   :1206.9

   medIncome      povertyPercent    MedianAge      MedianAgeMale
 Min.   : 22640   Min.   : 3.20   Min.   : 22.30   Min.   :22.40
 1st Qu.: 38883   1st Qu.:12.15   1st Qu.: 37.70   1st Qu.:36.35
 Median : 45207   Median :15.90   Median : 41.00   Median :39.60
 Mean   : 47063   Mean   :16.88   Mean   : 45.27   Mean   :39.57
 3rd Qu.: 52492   3rd Qu.:20.40   3rd Qu.: 44.00   3rd Qu.:42.50  
 Max.   :125635   Max.   :47.40   Max.   :624.00   Max.   :64.70  
 MedianAgeFemale AvgHouseholdSize PctMarriedHouseholds  PctNoHS18_24  
 Min.   :22.30   Min.   :1.86     Min.   :22.99        Min.   : 0.00  
 1st Qu.:39.10   1st Qu.:2.38     1st Qu.:47.76        1st Qu.:12.80  
 Median :42.40   Median :2.50     Median :51.67        Median :17.10  
 Mean   :42.15   Mean   :2.53     Mean   :51.24        Mean   :18.22  
 3rd Qu.:45.30   3rd Qu.:2.64     3rd Qu.:55.40        3rd Qu.:22.70  
 Max.   :65.70   Max.   :3.97     Max.   :78.08        Max.   :64.10  
   PctHS18_24   PctBachDeg18_24   PctHS25_Over   PctBachDeg25_Over
 Min.   : 0.0   Min.   : 0.000   Min.   : 7.50   Min.   : 2.50    
 1st Qu.:29.2   1st Qu.: 3.100   1st Qu.:30.40   1st Qu.: 9.40    
 Median :34.7   Median : 5.400   Median :35.30   Median :12.30    
 Mean   :35.0   Mean   : 6.158   Mean   :34.80   Mean   :13.28    
 3rd Qu.:40.7   3rd Qu.: 8.200   3rd Qu.:39.65   3rd Qu.:16.10    
 Max.   :72.5   Max.   :51.800   Max.   :54.80   Max.   :42.20    
 PctUnemployed16_Over PctPrivateCoverage PctEmpPrivCoverage
PctPublicCoverage
 Min.   : 0.400       Min.   :22.30      Min.   :13.5      
Min.   :11.20    
 1st Qu.: 5.500       1st Qu.:57.20      1st Qu.:34.5       1st
Qu.:30.90    
 Median : 7.600       Median :65.10      Median :41.1      
Median :36.30    
 Mean   : 7.852       Mean   :64.35      Mean   :41.2      
Mean   :36.25    
 3rd Qu.: 9.700       3rd Qu.:72.10      3rd Qu.:47.7       3rd
Qu.:41.55    
 Max.   :29.400       Max.   :92.30      Max.   :70.7      
Max.   :65.10    
 PctPublicCoverageAlone
 Min.   : 2.60         
 1st Qu.:14.85         
 Median :18.80         
 Mean   :19.24         
 3rd Qu.:23.10         
 Max.   :46.60         
```

Step 3: Create a regression model for death rate

```
# getting the names of all columns in our cleaned dataset
names(cancer_data_clean)
```

```
 [1] "County"                 "Population"              "deathRate"             

 [4] "incidenceRate"          "medIncome"              
"povertyPercent"        
 [7] "MedianAge"              "MedianAgeMale"          
"MedianAgeFemale"       
[10] "AvgHouseholdSize"       "PctMarriedHouseholds"   "PctNoHS18_24"          
```

```
[13] "PctHS18_24"              "PctBachDeg18_24"        "PctHS25_Over"

[16] "PctBachDeg25_Over"      "PctUnemployed16_Over"
"PctPrivateCoverage"
[19] "PctEmpPrivCoverage"     "PctPublicCoverage"
"PctPublicCoverageAlone"
```

```
# Create a regression model for death rate with existing "necessary"
columns
model_death_rate <- lm(deathRate ~ incidenceRate + medIncome +
povertyPercent + MedianAge + MedianAgeMale + MedianAgeFemale +
AvgHouseholdSize + PctMarriedHouseholds + PctNoHS18_24 + PctHS18_24 +
PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over +
PctUnemployed16_Over + PctPrivateCoverage + PctEmpPrivCoverage +
PctPublicCoverage + PctPublicCoverageAlone, data = cancer_data_clean)


# Summarize the model
summary(model_death_rate)
```

```
Call:
lm(formula = deathRate ~ incidenceRate + medIncome + povertyPercent +
    MedianAge + MedianAgeMale + MedianAgeFemale + AvgHouseholdSize +
    PctMarriedHouseholds + PctNoHS18_24 + PctHS18_24 + PctBachDeg18_24
+
    PctHS25_Over + PctBachDeg25_Over + PctUnemployed16_Over +
    PctPrivateCoverage + PctEmpPrivCoverage + PctPublicCoverage +
    PctPublicCoverageAlone, data = cancer_data_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-111.798  -10.745    0.047   10.573  142.257

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.450e+02  1.450e+01   9.998  < 2e-16 ***
incidenceRate           2.095e-01  6.587e-03  31.800  < 2e-16 ***
medIncome               1.065e-04  7.483e-05   1.424  0.15467
povertyPercent          5.667e-01  1.378e-01   4.112 4.03e-05 ***
MedianAge              -3.320e-03  7.702e-03  -0.431  0.66641
MedianAgeMale          -2.130e-01  1.971e-01  -1.081  0.27989
MedianAgeFemale        -2.386e-01  2.115e-01  -1.128  0.25953
AvgHouseholdSize       -1.622e+01  2.601e+00  -6.237 5.09e-10 ***
PctMarriedHouseholds   -5.351e-02  8.396e-02  -0.637  0.52392
PctNoHS18_24           -8.735e-02  5.435e-02  -1.607  0.10813
PctHS18_24              2.358e-01  4.819e-02   4.894 1.04e-06 ***
PctBachDeg18_24        -5.387e-02  1.051e-01  -0.512  0.60843
PctHS25_Over            5.129e-01  9.088e-02   5.643 1.83e-08 ***
PctBachDeg25_Over      -1.069e+00  1.462e-01  -7.310 3.40e-13 ***
```

```
PctUnemployed16_Over    6.123e-01  1.542e-01   3.970 7.36e-05 ***
PctPrivateCoverage     -5.756e-01  1.296e-01  -4.441 9.26e-06 ***
PctEmpPrivCoverage      2.906e-01  9.637e-02   3.016  0.00259 ** 
PctPublicCoverage      -1.392e-01  2.090e-01  -0.666  0.50532    
PctPublicCoverageAlone  1.824e-01  2.669e-01   0.683  0.49440    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.05 on 3028 degrees of freedom
Multiple R-squared:  0.5317,	Adjusted R-squared:  0.5289 
F-statistic:   191 on 18 and 3028 DF,  p-value: < 2.2e-16
```

Our Adjusted R-squared is around 0.5289. This is before we calculated the mortality rate

Calculating Mortality rate based on the incidenceRate and deathRate

To estimate the mortality rate, we need the number of deaths and the number of cancer cases. However, since we only have the death rate and the incidence rate (number of newly diagnosed cancer cases per 100,000 of population), we can estimate the mortality rate by multiplying the death rate and the inverse of the incidence rate.

```
# Therefore, We can estimate the mortality rate in our dataset as:
cancer_data_clean$Mortality_Rate <- cancer_data_clean$deathRate /
(1/cancer_data_clean$incidenceRate)

#We can now create our new regression model inclusive of the Moratlity
rate as follows:
model_death_rate_mr <- lm(deathRate ~ incidenceRate + medIncome +
povertyPercent + MedianAge + MedianAgeMale + MedianAgeFemale +
AvgHouseholdSize + PctMarriedHouseholds + PctNoHS18_24 + PctHS18_24 +
PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over +
PctUnemployed16_Over + PctPrivateCoverage + PctEmpPrivCoverage +
PctPublicCoverage + PctPublicCoverageAlone + Mortality_Rate, data =
cancer_data_clean)

# Summarize the new model (model_death_rate_mr)
summary(model_death_rate_mr)
```

```
Call:
lm(formula = deathRate ~ incidenceRate + medIncome + povertyPercent + 
    MedianAge + MedianAgeMale + MedianAgeFemale + AvgHouseholdSize + 
    PctMarriedHouseholds + PctNoHS18_24 + PctHS18_24 + PctBachDeg18_24 
+ 
    PctHS25_Over + PctBachDeg25_Over + PctUnemployed16_Over + 
    PctPrivateCoverage + PctEmpPrivCoverage + PctPublicCoverage + 
    PctPublicCoverageAlone + Mortality_Rate, data = cancer_data_clean)

Residuals:
     Min       1Q   Median       3Q      Max 
```

```
-255.933   -1.300    0.664    2.401   23.417

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             1.868e+02  5.129e+00  36.410  < 2e-16 ***
incidenceRate          -3.487e-01  4.481e-03 -77.824  < 2e-16 ***
medIncome               4.466e-05  2.643e-05   1.689 0.091245 .  
povertyPercent         -1.291e-01  4.891e-02  -2.640 0.008345 ** 
MedianAge              -1.882e-04  2.720e-03  -0.069 0.944858    
MedianAgeMale          -4.834e-01  6.963e-02  -6.943 4.69e-12 ***
MedianAgeFemale         2.517e-01  7.479e-02   3.366 0.000773 ***
AvgHouseholdSize       -1.956e+00  9.238e-01  -2.118 0.034289 *  
PctMarriedHouseholds   -7.189e-02  2.965e-02  -2.424 0.015393 *  
PctNoHS18_24           -7.718e-02  1.920e-02  -4.021 5.94e-05 ***
PctHS18_24             -2.124e-02  1.711e-02  -1.241 0.214591    
PctBachDeg18_24        -2.400e-02  3.714e-02  -0.646 0.518130    
PctHS25_Over            2.144e-01  3.217e-02   6.667 3.10e-11 ***
PctBachDeg25_Over      -3.023e-02  5.212e-02  -0.580 0.561969    
PctUnemployed16_Over    2.011e-01  5.455e-02   3.687 0.000231 ***
PctPrivateCoverage     -1.748e-01  4.586e-02  -3.811 0.000141 ***
PctEmpPrivCoverage      1.222e-01  3.406e-02   3.589 0.000337 ***
PctPublicCoverage       2.842e-01  7.387e-02   3.847 0.000122 ***
PctPublicCoverageAlone -2.809e-01  9.432e-02  -2.978 0.002924 ** 
Mortality_Rate          1.983e-03  1.361e-05 145.759  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.728 on 3027 degrees of freedom
Multiple R-squared:  0.9416,	Adjusted R-squared:  0.9412 
F-statistic:  2569 on 19 and 3027 DF,  p-value: < 2.2e-16
```

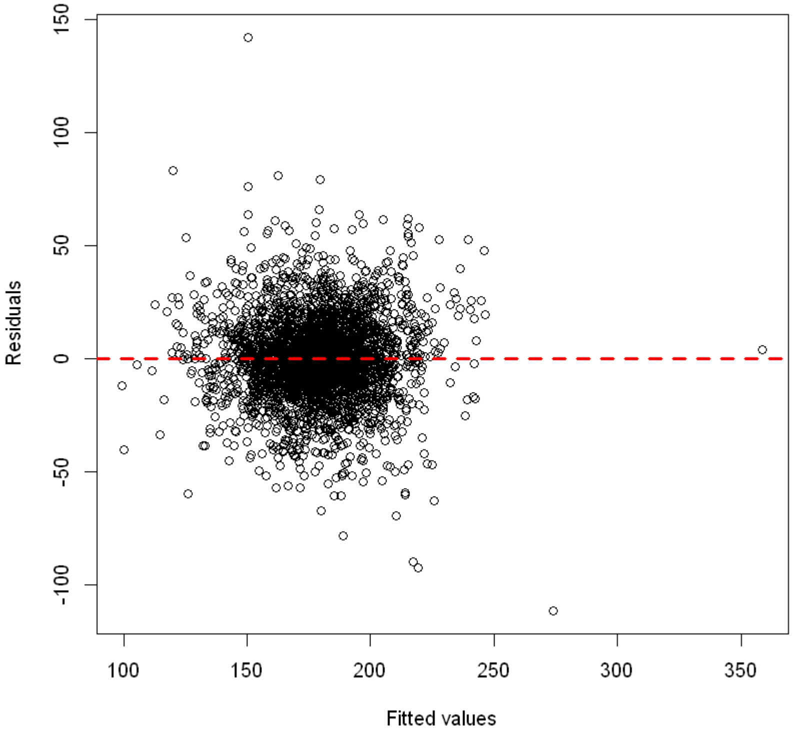
We now have a good score of 0.9412 on Adjusted R-squared and A p-value of < 2.2e-16 . This means Mortality rate as a predictor variable in our regression model is highly statistically significant.

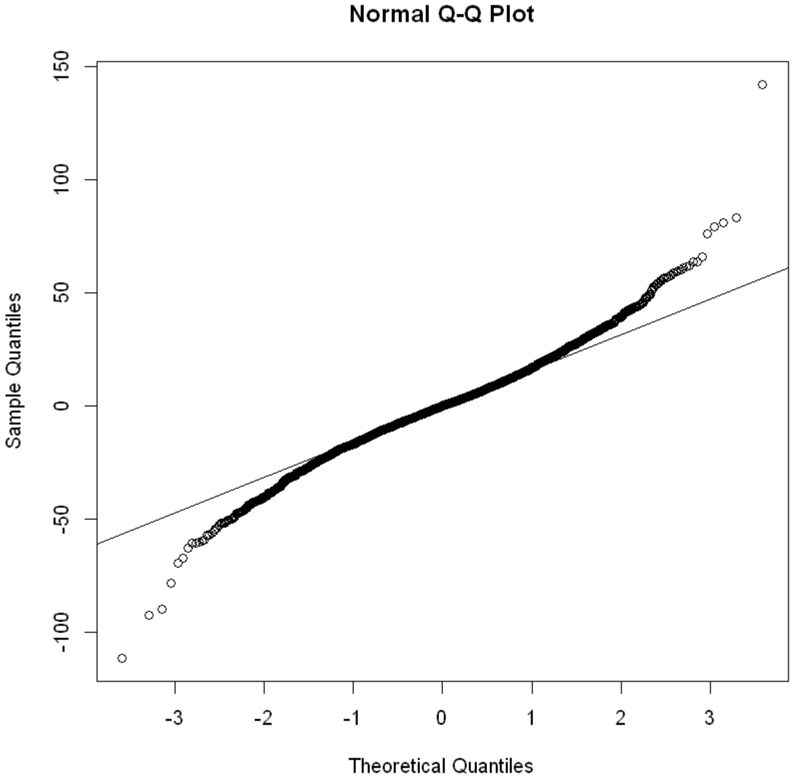
Step 4: Check the assumptions of the regression model

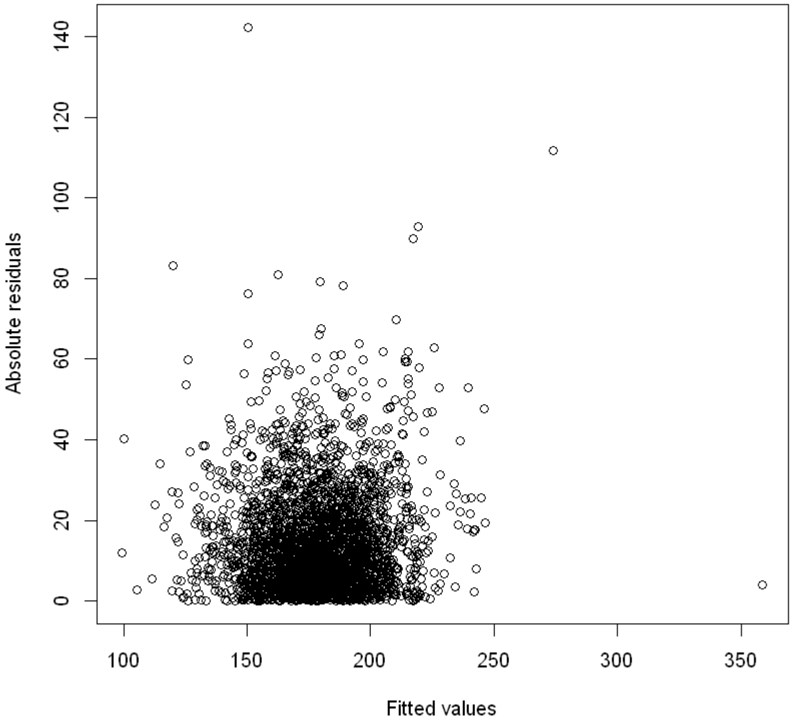
```
# Check the linearity assumption
plot(model_death_rate$fitted.values, model_death_rate$residuals, xlab
= "Fitted values", ylab = "Residuals")
abline(h = 0, lty = 2, col='red', lw=3)

# Check the normality assumption
qqnorm(model_death_rate$residuals)
qqline(model_death_rate$residuals)

# Check the homoscedasticity assumption
plot(model_death_rate$fitted.values, abs(model_death_rate$residuals),
xlab = "Fitted values", ylab = "Absolute residuals")
```


Residuals
Fitted values
150
100
50
0
-50
-100
100
150
200
250
300
350


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
150
100
50
0
-50
-100
-3
-2
-1
0
1
2
3


Absolute residuals
0
20
40
60
80
100
120
140
100
150
200
250
300
350
Fitted values