1. Introduction

The purpose of the explanatory data analysis which will be conducted in this report is to find an association between cancer mortality rate and demographic factors. The type of analysis can be helpful for future predictive studies and to inform authorities on distribution of healthcare resources. For this particular analysis, the dataset which will be used has been aggregated from a number of sources including the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov. The dataset includes information regarding the demographic statistics of different US counties and the associated cancer mortality rate.

The dataset contains 34 columns and 2000 rows of data. The response variable in the analysis is the cancer mortality rate, leaving a possible 33 explanatory variables. The dataset contains 33 numerical continuous variables and 1 categorical nominal variable (country location description). Below is a list of the variables used in the dataset:

- mortality: Mean *per capita* (100,000) cancer mortalities
- avgAnnCount: Mean number of reported cases of cancer diagnosed annually
- avgDeathsPerYear: Mean number of reported mortalities due to cancer
- incidenceRate: Mean *per capita* (100,000) cancer diagoses
- medianIncome: Median income per county
- popEst2015: Population of county
- povertyPercent: Percent of populace in poverty
- studyPerCap: *Per capita* number of cancer-related clinical trials per county
- binnedInc: Median income per capita binned by decile
- MedianAge: Median age of county residents
- MedianAgeMale: Median age of male county residents
- MedianAgeFemale: Median age of female county residents
- Geography: County name
- AvgHouseholdSize: Mean household size of county
- PercentMarried: Percent of county residents who are married
- PctNoHS18_24: Percent of county residents ages 18-24 highest education attained: less than high school
- PctHS18_24: Percent of county residents ages 18-24 highest education attained: high school diploma
- PctSomeCol18_24: Percent of county residents ages 18-24 highest education attained: some college
- PctBachDeg18_24: Percent of county residents ages 18-24 highest education attained: bachelor's degree
- PctHS25_Over: Percent of county residents ages 25 and over highest education attained: high school diploma
- PctBachDeg25_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree
- PctEmployed16_Over: Percent of county residents ages 16 and over employed
- PctUnemployed16_Over: Percent of county residents ages 16 and over unemployed
- PctPrivateCoverage: Percent of county residents with private health coverage

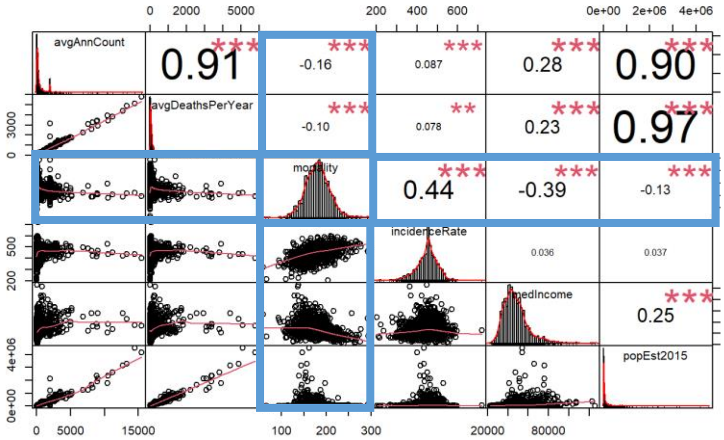
- PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance)
- PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage
- PctPublicCoverage: Percent of county residents with government-provided health coverage
- PctPubliceCoverageAlone: Percent of county residents with government-provided health coverage alone
- PctWhite: Percent of county residents who identify as White
- PctBlack: Percent of county residents who identify as Black
- PctAsian: Percent of county residents who identify as Asian
- PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian
- PctMarriedHouseholds: Percent of married households
- BirthRate: Number of live births relative to number of women in county

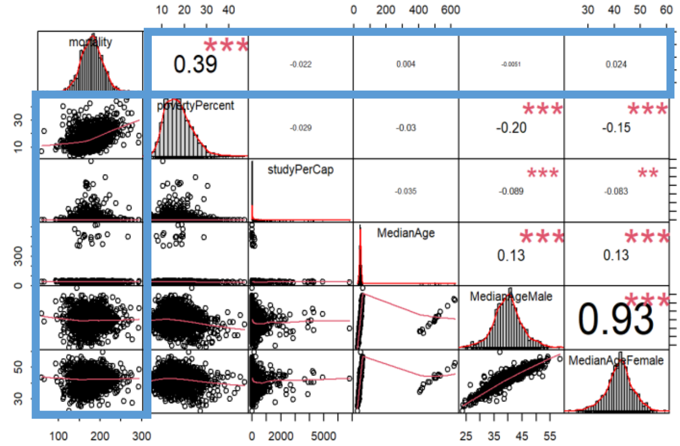
1. Exploratory Data Analysis

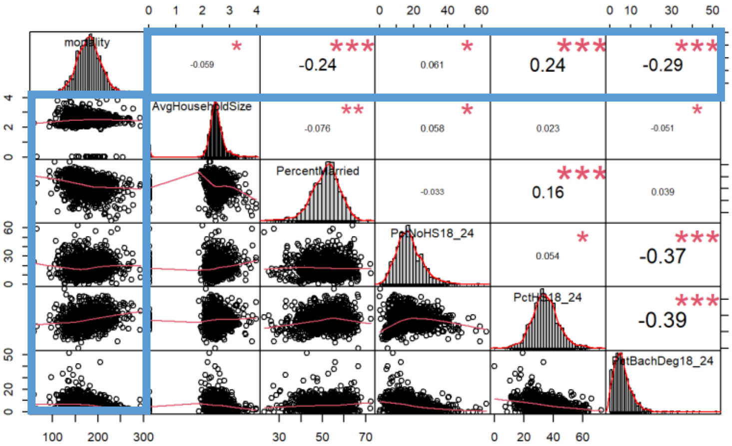
Before any explanatory data analysis can be performed the dataset needs to be transformed to be tidy. A new data frame was created which reflect the changes made to the original dataset. Firstly, the column labelled, *Geography* was removed since each row of the nominal this variable is different, hence it cannot be used as an explanatory variable. Another column which was removed labelled, *binnedInc*, since the data in this row was in the form of an ordered pair which was recognised in R as a string and not a continuous numerical value. Also, this column does not provide any additional useful information since the *medianIncome* column is the approximate average of the ordered pair. Additionally, the column titled, *PctSomeCol18_24* was removed since it contained too many missing values. Only 395 of the 2000 observations had values in the column *PctSomeCol18_24*.

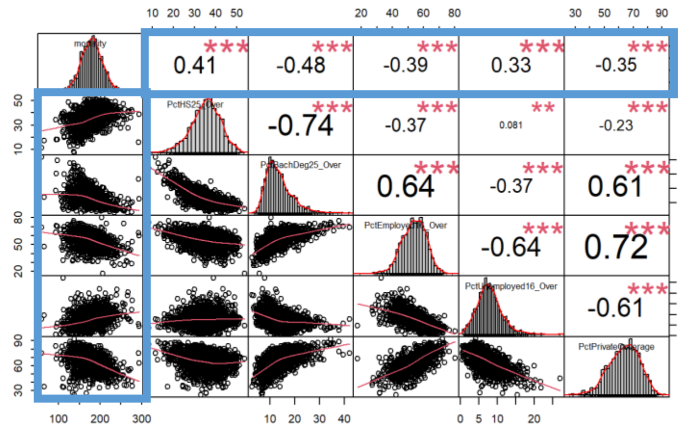
Finally, any rows which contained an empty value, 'N/A' were removed. After these rows were deleted, the new dataset contained 1534 observations and 31 variables. (See Appendix for R code).

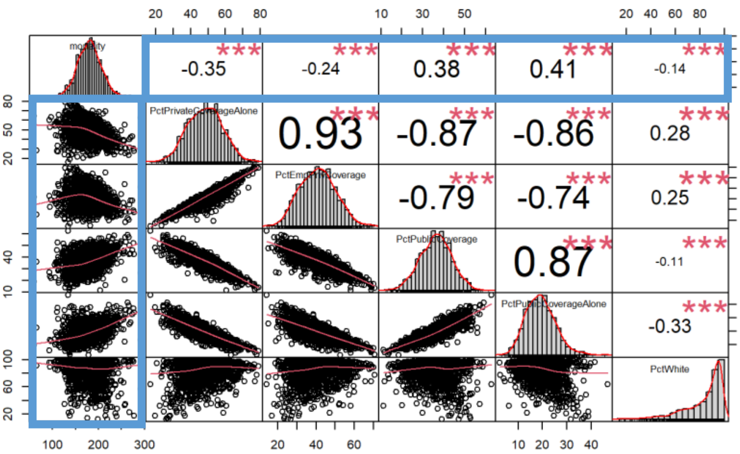
The correlation charts between cancer mortality rate and all the other remaining, 30 explanatory variables are shown below. The blue boxes which are added show the graphs and correlation coefficients which relate to cancer mortality rate (See Appendix for R code):

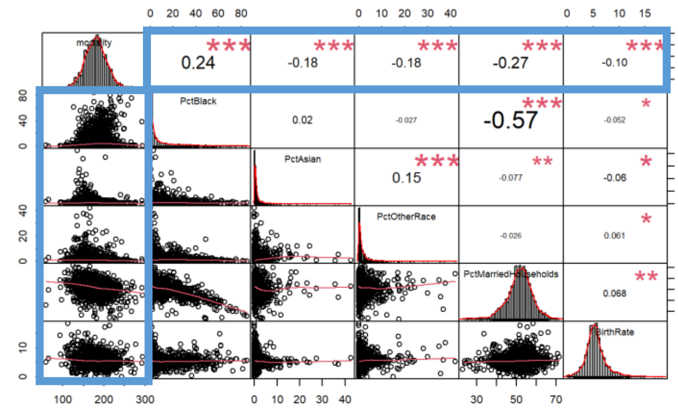






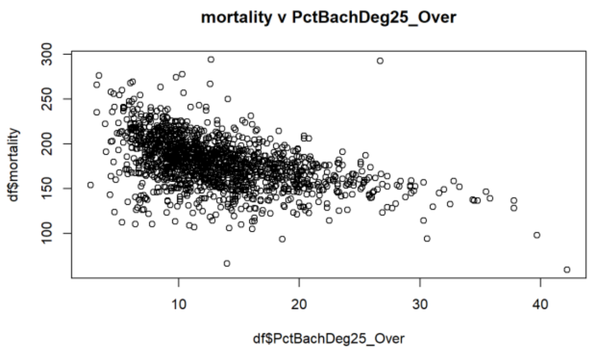


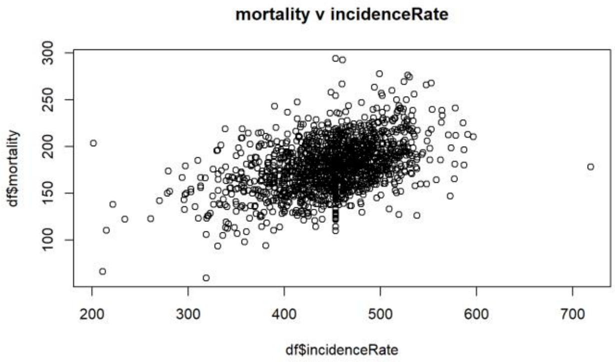


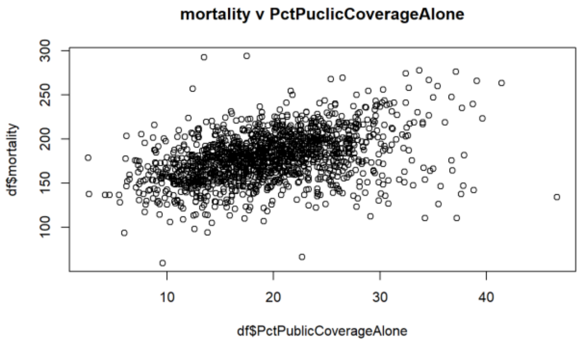


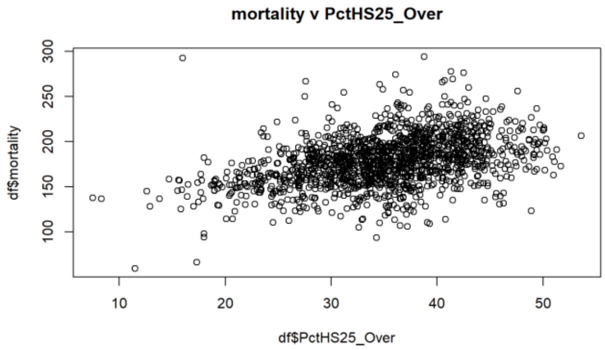
From the graphs above, the explanatory variables that yield the highest strength correlation (via Pearson's correlation coefficient) are: *PctBachDeg25_Over* (-0.48), *incidenceRate* (0.44), *PctPublicCoverageAlone* (0.41), *PctHS25_Over* (0.41). See Introduction for the description of these variable names.

The plots of the distribution of the cancer mortality rate against these four explanatory variables is shown below:



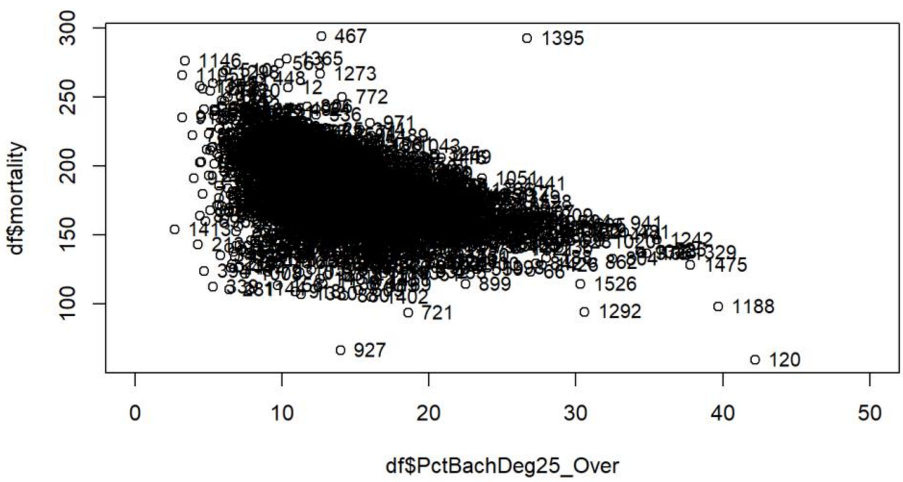




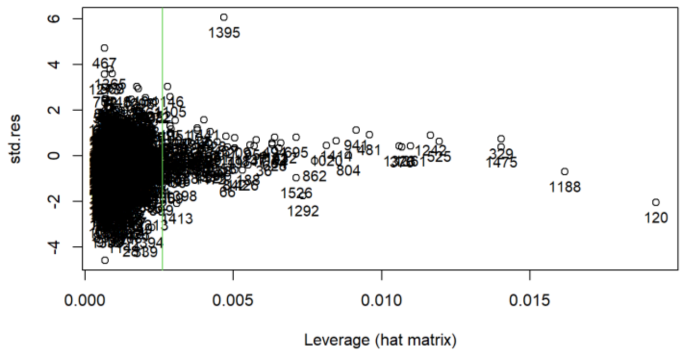


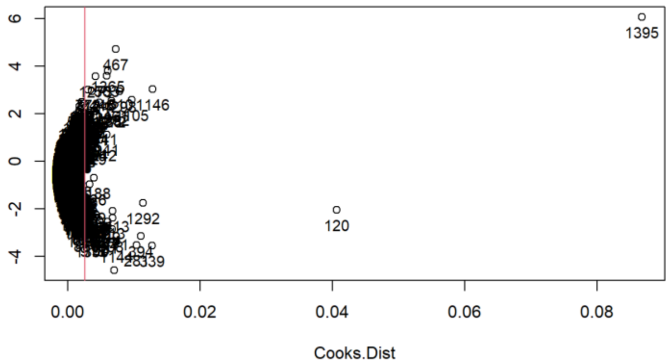
These four graphs above represent the four best simple linear regression models that could be fitted with cancer mortality rate as the response variable. However, it is possible that transformations of these graphs will produce better linear models. For example, in the graph above titled, *"mortality v PctBachDeg25_Over" (top right)*, the graph shows that the values are skewed to the right and the variance seems to decrease with an increase in x-values. Hence, a transformation could be applied to the x and y axis by using log scale for x-values and using a weighted regression.

Below is a graph of "*mortality v PctBachDeg25_Over",* with the points labelled.



From the graph above, a few outliers can be spotted such as: 1395 and 120, 927 etc. To find whether any of these points are influential, the graphs for leverage (hat matrix value) and Cook's distance must be graphed. Below both graphs are shown (See Appendix for R Code):



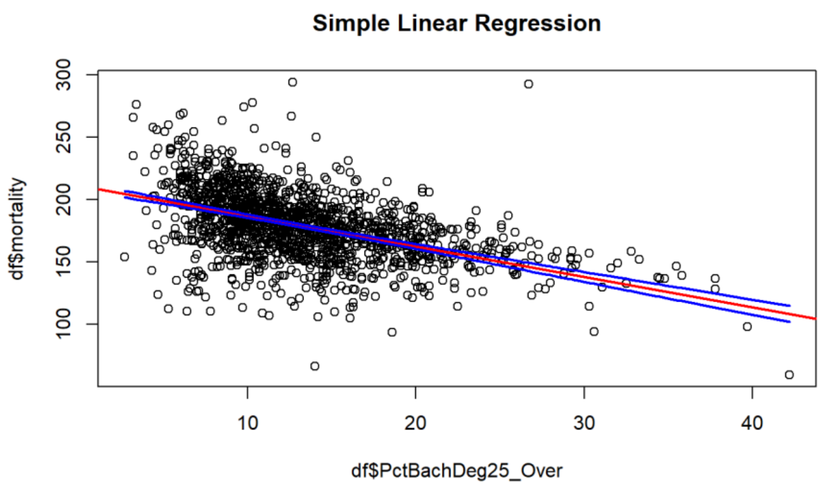


From the Cook's distance graph, the two points which show significant influence are: points 1395 and 120. Both points also exceed the hat matrix threshold, $h_i > \frac{4}{n}$. Since there are so many observations shown on the graph, these two points alone do not point to an alternative regression model and hence are candidates for removal from the model.

2. Simple Linear Regression: Analysis and Results

As shown in the previous section, the explanatory variable which best explains the response variable (cancer mortality rate) via a simple linear regression is: *PctBachDeg25_Over*. This variable represents the percent of county residents aged 25 and over whose highest attained education is a bachelor's degree.

A simple linear model of between *PctBachDeg25_Over* and *mortality* is shown below with the red line representing the regression and blue lines representing a 95% confidence interval (See Appendix for R code).



Below is the R output of the simple linear model:

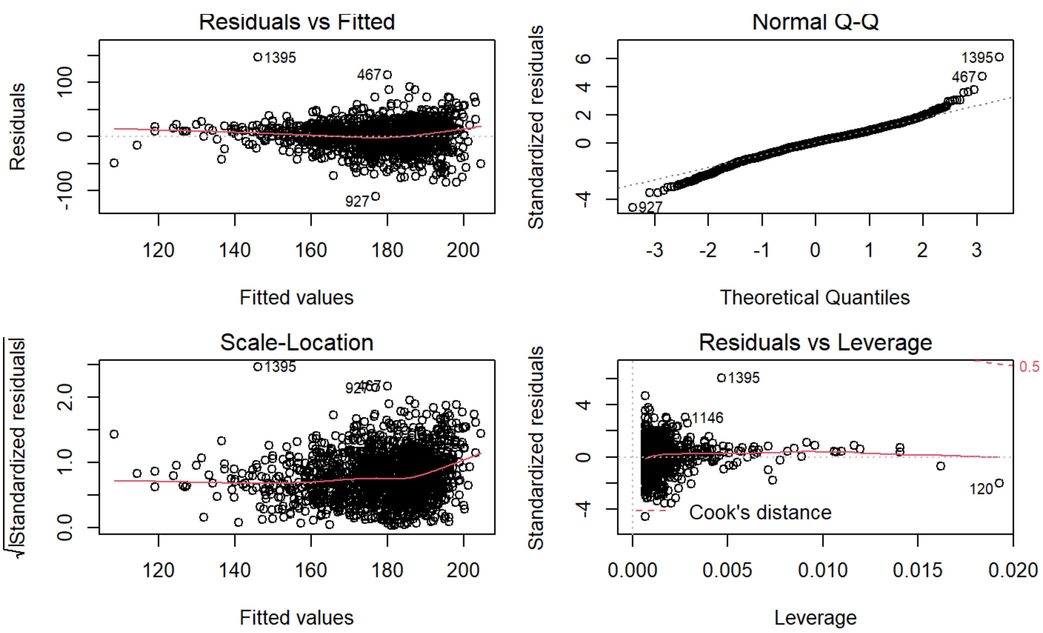
```
Call:
lm(formula = df$mortality ~ df$PctBachDeg25_Over)

Residuals:
     Min       1Q   Median       3Q      Max
-110.613  -14.356    1.133   14.301  146.442

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          210.9269     1.6227  129.98   <2e-16 ***
df$PctBachDeg25_Over  -2.4296     0.1137  -21.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.17 on 1532 degrees of freedom
Multiple R-squared:  0.2296,    Adjusted R-squared:  0.2291
F-statistic: 456.7 on 1 and 1532 DF,  p-value: < 2.2e-16
```

The simple regression model predicts that for every 1% increase in the population (>25 year age) having a bachelor's degree, results in a decrease the mortality rate of the county by 2.43 cancer related deaths per 100,000 people. The intercept of the simple linear model predicts that a county with a 0% of its population above 25 years old having a bachelor's degree would have a cancer mortality rate of 210.9 per 100,000 people. While this is the best predictive simple linear regression from the dataset, its adjusted coefficient of determination (R^2) is only: 0.2291. This means that only 22.91% of the variation in cancer mortality rate can be explained by the linear model (least squares line).

There are a few assumptions in the simple linear model which include:

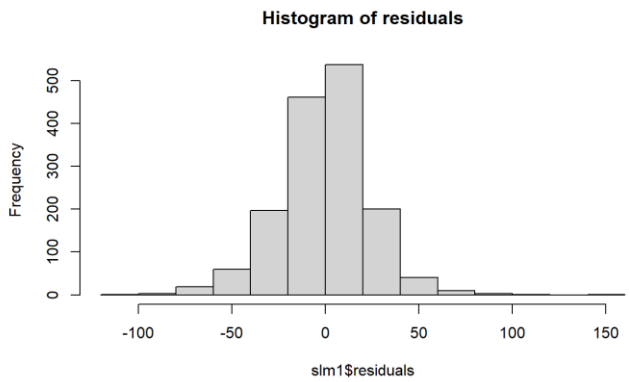
1) No influential outliers
2) Linearity (approximate model)
3) Independence of observations
4) Constant variance (homoscedasticity)
5) Errors (residuals) are random and normally distributed

Below is the application of simple diagnostic checking for these assumptions: (See Appendix for R code)



The first plot on the top left shows the residuals against the fitted values to check for randomness in the residuals (no pattern) and constant variance. The red smoothing curve is approximately flat and around the point 0, however is does increase slightly for fitted values > 190. This is also reflected in the standardised residuals curve (bottom right) where the smoothing curve tapers upwards for fitted values > 190. Hence the spread (variance) increases as x increases. Thus, the assumption for constant variance may not be satisfied.

The Q-Q plot shows that most observations lie around the straight line except at the ends where it starts to deviate. Overall, the deviations from the regression line (residuals) are approximately normally distributed. The histogram below also verifies that the residuals are approximately normal.



The bottom right plot of standardized residuals against leverage readily identifies any 'bad' leverage points. The cut-off Cook's distance used in the graph is 0.5, for which no observations exceed the bounds. However, if the cut-off used for Cook's distance is: 4/n, then there are several observations which exceed the bounds, with two that have a Cook's distance exceed 0.04. Although since these two observations do not exceed Cook's distance of 0.5, they aren't considered 'bad' leverage points.

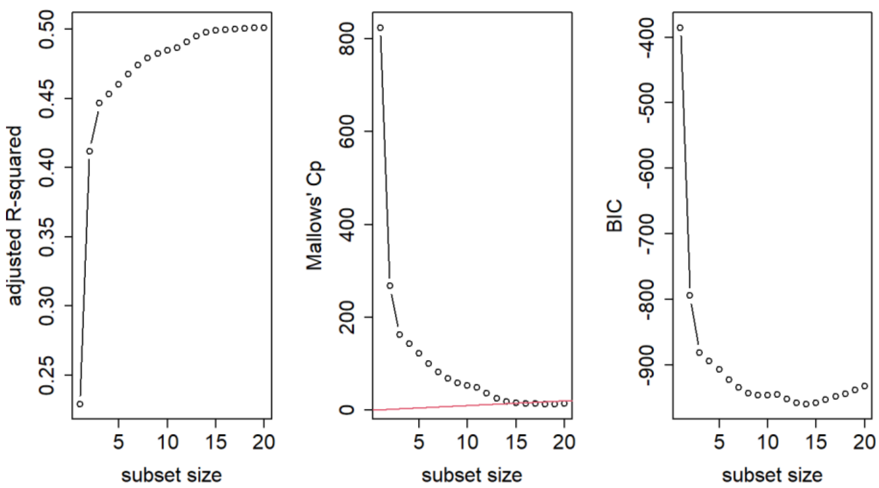
3. Multiple Linear Regression: Analysis and Results

For the one response variable of cancer mortality rate, there are a possible 30 explanatory variables in the dataset. This indicates that a multiple linear regression may produce a better model than a simple linear regression.

<u>All Subset Selection</u>

Both Mallow's C_p and BIC are two metrics used to pick the best regression model among several different models. The model is considered to be the best when both of these values are at a minimum. A similarity between the metrics is that the value of each metric will rise as more independent variables are included in the model. Statisticians much rather use a model containing less variables since, "If there are too many candidate variables, then the method fails to provide the best model, as some irrelevant variables are entered into the model" (National Library of Medicine, 2020).

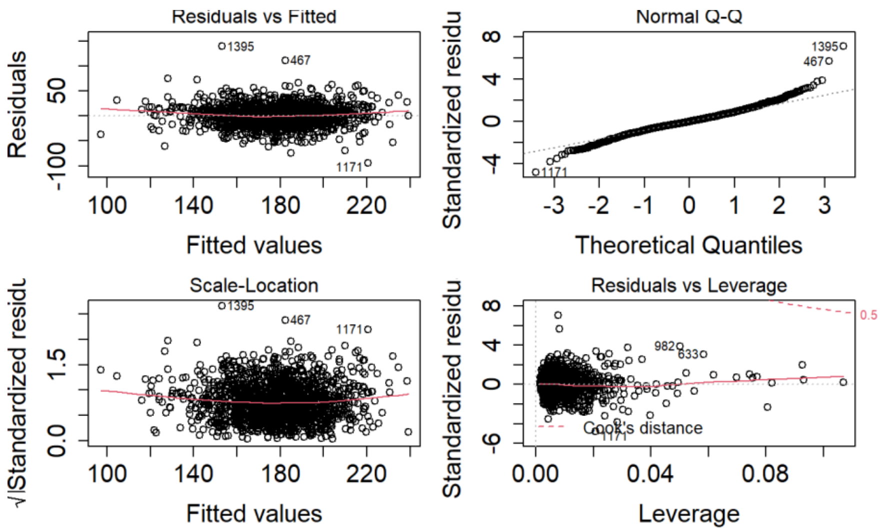
The first variable selection method that will be used for the multiple linear regression is: all subset selection. This all-subset selection was produced from selecting 1st best model for each subset size. The plots of R^2 , Mallows' C_p and BIC are shown below:



Good models ten to follow the property: $C_p \approx p$. Hence the points of intersection between the red line and the curve in the Mallow's C_p graph would give good sizes for candidate models.

Using the BIC and Mallow's C_p graph above, model evaluation was made on models consisting between 14 and 15 variables. (See Appendix for full summary of models). The adjusted R^2 for 14 variables was: 0.4839, while the adjusted R^2 for 15 variables was: 0.4865

The diagnostic plots for the 14 variable model are shown below:



The smoothing curve for the top left and bottom left graphs are slightly curved. Hence the assumption for constant variance may not be satisfied.

The Q-Q plot shows that most observations lie around the straight line except at the ends. Overall, the deviations from the regression line (residuals) are approximately normally distributed.

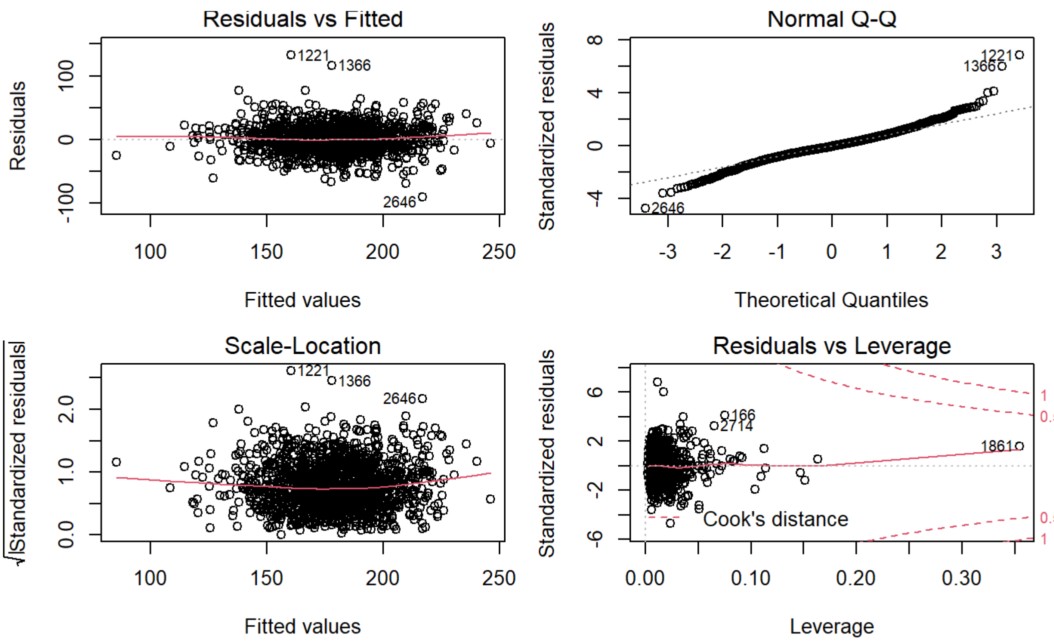
The bottom right plot of standardized residuals against leverage readily identifies any 'bad' leverage points. Using, the cut-off Cook's distance used in the graph as 0.5, there are no 'bad' leverage points in the model.

<u>Forward Selection</u>

The variable selection method used for the next multiple linear regression model is forward selection. The forward selection method starts with no variables included in the model and then it keeps fitting the next best variable. This process continues until the addition of an extra term increases the value of AIC (Akaike Information Criterion). AIC is a value which is maximised when the model can explain a great amount of variation, using the fewest independent variables (Scribbr, 2020). (See Appendix for full summary of model).

From this model, the variables which have the most influence on the forward selection model are: *incidenceRate* and *PctBachDeg25_Over* (since they have the lowest p-value). The adjusted R^2 value of this model is: 0.5009, meaning that 50.09% of the variation in cancer mortality rate can be explained by the multiple linear regression model.

The diagnostic plot of this forward selection model is shown below:



The first plot on the top left shows the residuals against the fitted values to check for randomness in the residuals (no pattern) and constant variance. The red smoothing curve is approximately flat and around the point 0. This is also reflected in the standardised residuals curve (bottom right) where the smoothing, has a slight curvature but is approximately flat. Thus, the assumption for constant variance and random errors is satisfied.

The Q-Q plot shows that most observations lie around the straight line except at the ends where it starts to deviate. Overall, the deviations from the regression line (residuals) are approximately normally distributed.

The bottom right plot of standardized residuals against leverage readily identifies any 'bad' leverage points. Using, the cut-off Cook's distance used in the graph as 0.5, there are no 'bad' leverage points in the model.

The final choice on which model to choose is the all-subset selection of 14 variables. This is because it uses a significantly smaller number of variables with an adjusted R^2 value only slightly less than the forward selection method.

4. Multivariate Regression

A multivariate regression has the advantage of using more than one response (dependant) variable within one model against multiple explanatory variables. A full model analysis was conducted with cancer mortality rate and *AvgAnnCount* (mean number of reported cancer cases diagnosed annually).
The final output of the regression between *AvgAnnCount* and all independent variables is: (See Appendix for R code and full output):

```
Residual standard error: 484.7 on 1503 degrees of freedom
Multiple R-squared:  0.8509,    Adjusted R-squared:  0.8479
F-statistic: 285.9 on 30 and 1503 DF,  p-value: < 2.2e-16
```

Using this new response variable for average annual count for cancer cases yields a much better adjusted R^2 value (0.8479) compared to cancer mortality rate. This indicates that there is a stronger association between cancer diagnosis with demographic factors rather than cancer deaths (per 100,000 people). However, this result must be treated with caution as annual diagnosis does not account for population size of the county. Population of the county is a large factor since many of the explanatory variables are as percentages.

5. Discussion/Conclusion

Through the data analysis, associations were found between the cancer mortality and different predictor variables. From the simple linear regression, the variable that had the greatest association with the response variable was: *PctBachDeg25_Over* (Percent of county residents ages 25 and over highest education attained: bachelor's degree). Despite having the largest correlation coefficient, only 22.91% of the variation in the response variable could be explained by the model.

When a multiple linear regression was applied, the chosen model consisted of 14 independent variables which was selected from all subset selection method. This model could explain 48.39% of the variation in the response variable. Hence, the multiple linear regression is the preferred model using adjusted R^2 as the criterion.

There are limitations to this study as fitting a linear regression requires the fulfillment of several factors. It is unknown whether the data was collected randomly and independently, which is an assumption for any linear regression model. There is also a lack of information in the dataset of over how long this data has been culminated. Further data exploration could be conducted, given the date of the data input. This information would provide for a better understanding of how cancer mortality rate in changing in different US counties.

6. References

Bevans, R. (2020, March 26). *Akaike Information Criterion | When & How to Use It*. Scribbr. https://www.scribbr.com/statistics/akaike-information-criterion/

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), e000262. https://doi.org/10.1136/fmch-2019-000262

7. R code and Output

R code:

```
#Exploratory Data Analysis
print(load("cancer.RData"))
library(PerformanceAnalytics)
library("magrittr")
cancer_project <- cancer_project[-c(18,13,9)]
df <- na.omit(cancer_project)
df$row <- c(1:1534)

df1 <- df[1:6]
df2 <- df[c(3,7:11)]
df3 <- df[c(3,12:16)]
df4 <- df[c(3,17:21)]
df5 <- df[c(3,22:26)]
df6 <- df[c(3,27:31)]

pairs(df1)
chart.Correlation(df1)
pairs(df2)
chart.Correlation(df2)
pairs(df3)
chart.Correlation(df3)
pairs(df4)
chart.Correlation(df4)
pairs(df5)
chart.Correlation(df5)
pairs(df6)
chart.Correlation(df6)

plot(df$mortality ~ df$PctBachDeg25_Over, xlim=c(0,50))
text(df$PctBachDeg25_Over,df$mortality, labels = df$row, pos = 4, cex=0.9)
plot(df$mortality ~ df$incidenceRate)
plot(df$mortality ~ df$PctPublicCoverageAlone)
plot(df$mortality ~ df$PctHS25_Over)
```

```
#Simple Linear Model
slm1 <- lm(df$mortality ~ df$PctBachDeg25_Over)
summary(slm1)
conf.mean=predict(slm1,interval="confidence",level=0.95)
plot(df$mortality ~ df$PctBachDeg25_Over, main = "Simple Linear Regression")
abline(slm1, col = "red", lwd =2)
matlines(sort(df$PctBachDeg25_Over),
    conf.mean[order(df$PctBachDeg25_Over), 2:3],
    lwd = 2, col = "blue",
            lty = 1)

plot(slm1)
res=slm1$residuals
std.res=rstandard(slm1)  ## standardised residuals

Leverage<-hatvalues(slm1)
tail(sort(Leverage))

Cooks.Dist<-cooks.distance(slm1)
tail(sort(Cooks.Dist))

p=1
n=1534
plot(std.res~Leverage, xlab = "Leverage (hat matrix)")
text(Leverage,std.res, labels = df$row, pos = 1, cex=0.9)
abline(v=2*(p+1)/n, lty=1,col=3)  ## add a vertical line for the cut-off Leverage
plot(std.res~Cooks.Dist)
text(Cooks.Dist,std.res, labels = df$row, pos = 1, cex=0.9, xlim=c(0,2))
abline(v=2*(p+1)/(n-(p+1)), lty=1, col=2)  ## add a vertical line for the cut-off Cooks

plot(cooks.distance(slm1), xlab = "Locations", ylab = "Cook's distance")
abline(h=2*(p+1)/(n-(p+1)), lty=1, col=2)
with(df, text(cooks.distance(slm1), labels = df$row, pos = 4))

par(mfrow=c(2,2))
par(mar = c(5, 4, 1, 1) + 0.1) # This isn't necessary, but gives better margin spacing around the plot
plot(slm1)

hist(slm1$residuals, main = "Histogram of residuals")

#Multiple Linear Model
par(mfrow = c(1, 3))
par(cex.axis = 1.5)
par(cex.lab = 1.5)
AllSubsets <- regsubsets(mortality ~ ., nvmax = 20, nbest = 1, data = df)
AllSubsets.summary <- summary(AllSubsets)
plot(1:20, AllSubsets.summary$adjr2, xlab = "subset size", ylab = "adjusted R-squared", type = "b")
plot(1:20, AllSubsets.summary$cp, xlab = "subset size", ylab = "Mallows' Cp", type = "b")
abline(0,1,col=2)
plot(1:20, AllSubsets.summary$bic, xlab = "subset size", ylab = "BIC", type = "b")

summary(AllSubsets)

lm14 <- lm(df$mortality ~ df$avgAnnCount + df$incidenceRate + df$MedianAgeFemale + df$PercentMarried +
df$PctNoHS18_24 + df$PctHS25_Over + df$PctBachDeg25_Over + df$PctEmployed16_Over + df$PctPrivateCoverage +
df$PctMarriedHouseholds + df$PctOtherRace + df$BirthRate)

lm15 <- lm(df$mortality ~ df$avgDeathsPerYear + df$avgAnnCount + df$incidenceRate + df$MedianAgeFemale +
df$PercentMarried + df$PctNoHS18_24 + df$PctHS25_Over + df$PctBachDeg25_Over + df$PctEmployed16_Over +
df$PctPrivateCoverage + df$PctMarriedHouseholds + df$PctOtherRace + df$BirthRate)
```

```
summary(lm14)
summary(lm15)
par(mfrow=c(2,2))
par(mar = c(5, 4, 1, 1) + 0.1)
plot(lm14)

res=lm.forward$residuals
std.res=rstandard(lm.forward)  ## standardised residuals
par(mfrow=c(2,2))## plotting 3 plots to check normality and constant variance
par(mar = c(5, 4, 1, 1) + 0.1)
qqnorm(std.res)
qqline(std.res)
plot(std.res,xlab="Time", ylab="Standardised Residuals")
plot(lm.forward$fitted.values,std.res, xlab="Fitted Values", ylab="Standardised Residuals")
par(mfrow=c(1,1))


Leverage<-hatvalues(lm.forward)
tail(sort(Leverage))

Cooks.Dist<-cooks.distance(lm.forward)
tail(sort(Cooks.Dist))

p=22
n=1534
par(mfrow=c(1,2))
par(mgp=c(1.75,0.75,0))
par(mar=c(3,3,2,1))
plot(std.res~Leverage)
abline(v=2*(p+1)/n, lty=1,col=3)  ## add a vertical line for the cut-off Leverage
plot(std.res~Cooks.Dist)
abline(v=2*(p+1)/(n-(p+1)), lty=1, col=2)  ## add a vertical line for the cut-off Cooks

plot(cooks.distance(lm.forward), xlab = "Locations", ylab = "Cook's distance")
abline(h=2*(p+1)/(n-(p+1)), lty=1, col=2)
with(df, text(cooks.distance(lm.forward), labels = row.names(df), pos = 4))

#Multivariate Linear Model
mlm1 <- lm(cbind(mortality, incidenceRate) ~ ., data = df)
summary(mlm1)
```

R output – Not in body of report

Summary of 14 variable linear model:

```
Call:
lm(formula = df$mortality ~ df$avgAnnCount + df$incidenceRate + 
    df$MedianAgeFemale + df$PercentMarried + df$PctNoHS18_24 + 
    df$PctHS25_Over + df$PctBachDeg25_Over + df$PctEmployed16_Over + 
    df$PctPrivateCoverage + df$PctMarriedHouseholds + df$PctOtherRace + 
    df$BirthRate)

Residuals:
    Min      1Q  Median      3Q     Max 
-94.079 -10.918  -0.637  11.149 139.253 

Coefficients:
```

```
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.966e+02  1.019e+01  19.292  < 2e-16 ***
df$avgAnnCount          -1.000e-03  4.515e-04  -2.215 0.026929 *  
df$incidenceRate         2.090e-01  1.062e-02  19.672  < 2e-16 ***
df$MedianAgeFemale      -9.434e-01  1.542e-01  -6.117 1.21e-09 ***
df$PercentMarried        7.379e-01  2.146e-01   3.438 0.000601 ***
df$PctNoHS18_24         -2.233e-01  7.373e-02  -3.028 0.002501 ** 
df$PctHS25_Over          5.180e-01  1.299e-01   3.987 7.01e-05 ***
df$PctBachDeg25_Over    -1.219e+00  2.039e-01  -5.977 2.82e-09 ***
df$PctEmployed16_Over   -4.541e-01  1.241e-01  -3.660 0.000261 ***
df$PctPrivateCoverage   -3.912e-01  9.158e-02  -4.271 2.06e-05 ***
df$PctMarriedHouseholds -9.653e-01  1.943e-01  -4.968 7.51e-07 ***
df$PctOtherRace         -6.293e-01  1.577e-01  -3.991 6.89e-05 ***
df$BirthRate            -1.216e+00  2.618e-01  -4.646 3.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.78 on 1521 degrees of freedom
Multiple R-squared:  0.488,	Adjusted R-squared:  0.4839 
F-statistic: 120.8 on 12 and 1521 DF,  p-value: < 2.2e-16
```

Summary of 15 variable model:

```
Call:
lm(formula = df$mortality ~ df$avgDeathsPerYear + df$avgAnnCount + 
    df$incidenceRate + df$MedianAgeFemale + df$PercentMarried + 
    df$PctNoHS18_24 + df$PctHS25_Over + df$PctBachDeg25_Over + 
    df$PctEmployed16_Over + df$PctPrivateCoverage + df$PctMarriedHouseholds + 
    df$PctOtherRace + df$BirthRate)

Residuals:
    Min      1Q  Median      3Q     Max 
-93.334 -10.779  -0.752  11.084 138.268 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             192.571769  10.260871  18.768  < 2e-16 ***
df$avgDeathsPerYear       0.008672   0.002949   2.940 0.003327 ** 
df$avgAnnCount           -0.003716   0.001028  -3.616 0.000309 ***
df$incidenceRate          0.209042   0.010597  19.727  < 2e-16 ***
df$MedianAgeFemale       -0.914767   0.154151  -5.934 3.65e-09 ***
df$PercentMarried         0.775377   0.214455   3.616 0.000309 ***
df$PctNoHS18_24          -0.207259   0.073745  -2.811 0.005010 ** 
df$PctHS25_Over           0.485698   0.130073   3.734 0.000195 ***
df$PctBachDeg25_Over     -1.308306   0.205689  -6.361 2.65e-10 ***
df$PctEmployed16_Over    -0.428439   0.124075  -3.453 0.000569 ***
df$PctPrivateCoverage    -0.359737   0.091976  -3.911 9.59e-05 ***
df$PctMarriedHouseholds  -0.982085   0.193886  -5.065 4.58e-07 ***
df$PctOtherRace          -0.643756   0.157366  -4.091 4.52e-05 ***
df$BirthRate             -1.136110   0.262584  -4.327 1.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.73 on 1520 degrees of freedom
Multiple R-squared:  0.4909,	Adjusted R-squared:  0.4865 
F-statistic: 112.7 on 13 and 1520 DF,  p-value: < 2.2e-16
```

Summary of forward selection model:

```
Call:
lm(formula = mortality ~ PctBachDeg25_Over + incidenceRate + 
    povertyPercent + PctOtherRace + PctWhite + BirthRate + PctHS18_24 + 
    MedianAgeFemale + PctHS25_Over + PctPrivateCoverage + PctNoHS18_24 + 
    PctEmpPrivCoverage + avgAnnCount + avgDeathsPerYear + PctMarriedHouseholds + 
    PercentMarried + PctEmployed16_Over + medIncome + popEst2015, 
    data = df)

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

```
-90.424 -10.494  -0.598  10.439 132.378

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.692e+02  1.891e+01   8.947  < 2e-16 ***
PctBachDeg25_Over    -1.462e+00  2.134e-01  -6.853 1.05e-11 ***
incidenceRate         1.993e-01  1.063e-02  18.753  < 2e-16 ***
povertyPercent        4.272e-01  2.287e-01   1.868 0.061988 .  
PctOtherRace         -6.971e-01  1.580e-01  -4.412 1.10e-05 ***
PctWhite             -1.680e-01  4.725e-02  -3.555 0.000390 ***
BirthRate            -1.103e+00  2.635e-01  -4.185 3.01e-05 ***
PctHS18_24            9.760e-02  6.863e-02   1.422 0.155173    
MedianAgeFemale      -6.675e-01  1.689e-01  -3.953 8.08e-05 ***
PctHS25_Over          4.400e-01  1.361e-01   3.234 0.001249 ** 
PctPrivateCoverage   -5.774e-01  1.313e-01  -4.397 1.17e-05 ***
PctNoHS18_24         -2.483e-01  7.605e-02  -3.265 0.001118 ** 
PctEmpPrivCoverage    4.748e-01  1.343e-01   3.536 0.000419 ***
avgAnnCount          -3.182e-03  1.025e-03  -3.104 0.001946 ** 
avgDeathsPerYear      1.468e-02  5.653e-03   2.598 0.009475 ** 
PctMarriedHouseholds -1.263e+00  2.147e-01  -5.884 4.93e-09 ***
PercentMarried        1.323e+00  2.307e-01   5.733 1.19e-08 ***
PctEmployed16_Over   -5.232e-01  1.408e-01  -3.717 0.000209 ***
medIncome             2.233e-04  1.083e-04   2.062 0.039384 *  
popEst2015           -1.352e-05  7.968e-06  -1.697 0.089869 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.45 on 1514 degrees of freedom
Multiple R-squared:  0.507,	Adjusted R-squared:  0.5009 
F-statistic: 81.96 on 19 and 1514 DF,  p-value: < 2.2e-16
```

Summary of multivariate model:

```
Response mortality :

Call:
lm(formula = mortality ~ avgDeathsPerYear + incidenceRate + medIncome + 
    popEst2015 + povertyPercent + studyPerCap + MedianAge + MedianAgeMale + 
    MedianAgeFemale + AvgHouseholdSize + PercentMarried + PctNoHS18_24 + 
    PctHS18_24 + PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over + 
    PctEmployed16_Over + PctUnemployed16_Over + PctPrivateCoverage + 
    PctPrivateCoverageAlone + PctEmpPrivCoverage + PctPublicCoverage + 
    PctPublicCoverageAlone + PctWhite + PctBlack + PctAsian + 
    PctOtherRace + PctMarriedHouseholds + BirthRate + row, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-91.172 -10.637  -0.633  10.362 133.184 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.871e+02  2.265e+01   8.260 3.15e-16 ***
avgDeathsPerYear         9.173e-03  5.348e-03   1.715 0.086527 .  
incidenceRate            1.988e-01  1.097e-02  18.123  < 2e-16 ***
medIncome                2.670e-04  1.150e-04   2.323 0.020320 *  
popEst2015              -1.683e-05  8.089e-06  -2.081 0.037607 *  
povertyPercent           4.816e-01  2.378e-01   2.025 0.043004 *  
studyPerCap             -3.360e-04  1.093e-03  -0.307 0.758510    
MedianAge               -3.015e-03  1.008e-02  -0.299 0.764908    
MedianAgeMale           -5.002e-02  2.899e-01  -0.173 0.863031    
MedianAgeFemale         -6.405e-01  3.161e-01  -2.026 0.042892 *  
AvgHouseholdSize         7.172e-01  1.319e+00   0.544 0.586719    
PercentMarried           1.367e+00  2.421e-01   5.649 1.93e-08 ***
PctNoHS18_24            -3.025e-01  7.977e-02  -3.792 0.000155 ***
PctHS18_24               9.225e-02  7.216e-02   1.278 0.201341    
PctBachDeg18_24         -1.749e-01  1.525e-01  -1.147 0.251568    
PctHS25_Over             4.785e-01  1.397e-01   3.425 0.000630 ***
PctBachDeg25_Over       -1.372e+00  2.247e-01  -6.107 1.29e-09 ***
PctEmployed16_Over      -6.273e-01  1.578e-01  -3.977 7.32e-05 ***
PctUnemployed16_Over    -2.693e-01  2.463e-01  -1.093 0.274496    
PctPrivateCoverage      -6.524e-01  3.701e-01  -1.763 0.078168 .  
PctPrivateCoverageAlone -1.065e-01  4.641e-01  -0.230 0.818476    
PctEmpPrivCoverage       5.447e-01  1.867e-01   2.918 0.003579 ** 
```

```
PctPublicCoverage       -5.975e-02  4.553e-01  -0.131 0.895611    
PctPublicCoverageAlone  -1.319e-01  5.099e-01  -0.259 0.795909    
PctWhite                -1.607e-01  8.491e-02  -1.892 0.058623 .  
PctBlack                 1.777e-02  8.382e-02   0.212 0.832162    
PctAsian                -1.441e-01  2.574e-01  -0.560 0.575614    
PctOtherRace            -7.091e-01  1.705e-01  -4.159 3.38e-05 ***
PctMarriedHouseholds    -1.353e+00  2.317e-01  -5.838 6.46e-09 ***
BirthRate               -1.206e+00  2.667e-01  -4.521 6.65e-06 ***
row                     -5.491e-04  1.137e-03  -0.483 0.629342    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.55 on 1503 degrees of freedom
Multiple R-squared:  0.5056,	Adjusted R-squared:  0.4957 
F-statistic: 51.23 on 30 and 1503 DF,  p-value: < 2.2e-16


Response avgAnnCount :

Call:
lm(formula = avgAnnCount ~ avgDeathsPerYear + incidenceRate + 
    medIncome + popEst2015 + povertyPercent + studyPerCap + MedianAge + 
    MedianAgeMale + MedianAgeFemale + AvgHouseholdSize + PercentMarried + 
    PctNoHS18_24 + PctHS18_24 + PctBachDeg18_24 + PctHS25_Over + 
    PctBachDeg25_Over + PctEmployed16_Over + PctUnemployed16_Over + 
    PctPrivateCoverage + PctPrivateCoverageAlone + PctEmpPrivCoverage + 
    PctPublicCoverage + PctPublicCoverageAlone + PctWhite + PctBlack + 
    PctAsian + PctOtherRace + PctMarriedHouseholds + BirthRate + 
    row, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-6304.0  -199.9   -93.3    39.0  2143.5 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -1.861e+03  5.615e+02  -3.315 0.000939 ***
avgDeathsPerYear         2.018e+00  1.326e-01  15.219  < 2e-16 ***
incidenceRate            7.833e-01  2.719e-01   2.880 0.004028 ** 
medIncome                1.963e-03  2.850e-03   0.689 0.491078    
popEst2015               8.710e-04  2.005e-04   4.344 1.49e-05 ***
povertyPercent           3.903e+00  5.894e+00   0.662 0.507959    
studyPerCap              7.328e-02  2.709e-02   2.705 0.006903 ** 
MedianAge               -4.966e-03  2.499e-01  -0.020 0.984147    
MedianAgeMale           -4.142e+00  7.186e+00  -0.576 0.564432    
MedianAgeFemale          1.351e+01  7.835e+00   1.725 0.084762 .  
AvgHouseholdSize        -9.687e+00  3.270e+01  -0.296 0.767070    
PercentMarried           2.273e+00  6.001e+00   0.379 0.704888    
PctNoHS18_24             4.150e+00  1.977e+00   2.099 0.036018 *  
PctHS18_24              -5.398e+00  1.789e+00  -3.018 0.002589 ** 
PctBachDeg18_24          5.779e-01  3.779e+00   0.153 0.878490    
PctHS25_Over            -1.045e+01  3.463e+00  -3.019 0.002582 ** 
PctBachDeg25_Over       -1.847e+01  5.569e+00  -3.316 0.000936 ***
PctEmployed16_Over       1.414e+01  3.910e+00   3.615 0.000310 ***
PctUnemployed16_Over    -2.664e+00  6.106e+00  -0.436 0.662734    
PctPrivateCoverage       4.954e+00  9.175e+00   0.540 0.589334    
PctPrivateCoverageAlone  1.618e+01  1.151e+01   1.406 0.159971    
PctEmpPrivCoverage      -8.917e+00  4.628e+00  -1.927 0.054198 .  
PctPublicCoverage        1.886e+01  1.129e+01   1.671 0.094916 .  
PctPublicCoverageAlone  -6.651e+00  1.264e+01  -0.526 0.598784    
PctWhite                -4.402e+00  2.105e+00  -2.091 0.036660 *  
PctBlack                -5.400e+00  2.078e+00  -2.599 0.009451 ** 
PctAsian                -3.007e-01  6.380e+00  -0.047 0.962418    
PctOtherRace             2.583e+00  4.227e+00   0.611 0.541247    
PctMarriedHouseholds    -3.964e-01  5.743e+00  -0.069 0.944987    
BirthRate                2.265e+01  6.612e+00   3.425 0.000631 ***
row                     -2.827e-02  2.819e-02  -1.003 0.316147    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 484.7 on 1503 degrees of freedom
Multiple R-squared:  0.8509,	Adjusted R-squared:  0.8479 
F-statistic: 285.9 on 30 and 1503 DF,  p-value: < 2.2e-16
```