Group Project

Andrea, Ben, Dustin

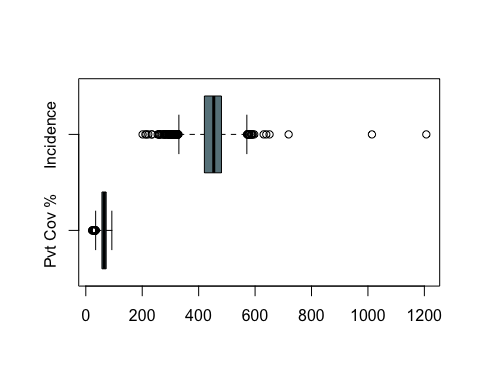
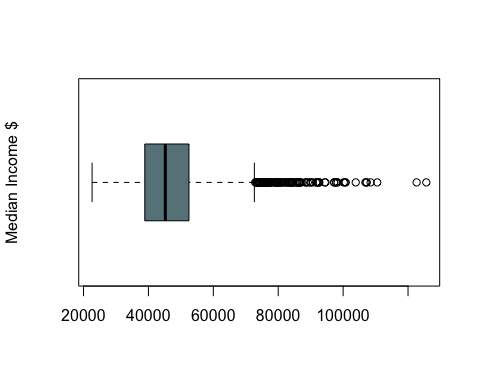
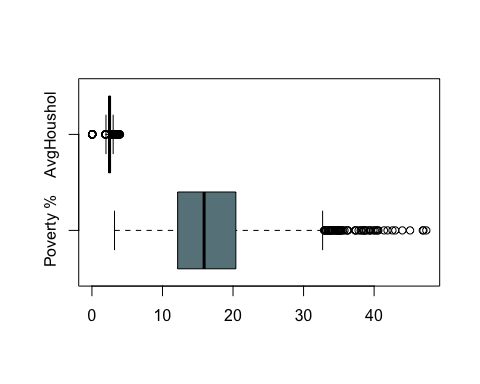
10/6/2021

### Intro

For this project we selected a dataset with dimension 3047x5. It relates the death rate of cancer in the United State divided in 4 macro areas (our categorical) with 5 quantitative predictors: Percentage of Private Coverage, Incidence Rate, Median Income, Average Household, and Poverty Percent.

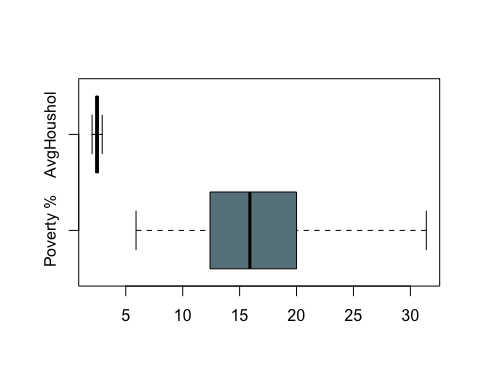
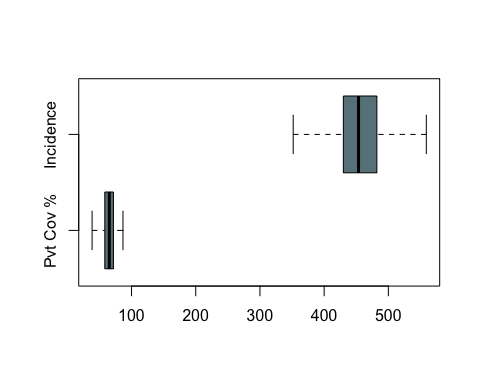
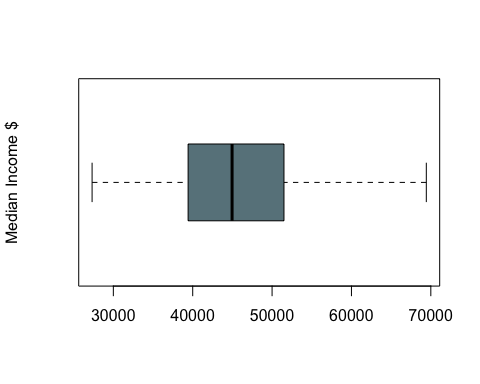
### Scatterplot analysis

From the scatterplots of the 5 predictor variables, it can be noticed that are presents numerous outliers in the dataset. Therefore we ran a routine in R that eliminated those outliers.



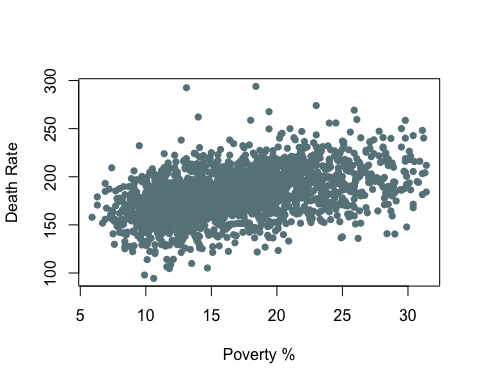
### Results

It can be noticed that the routine in R eliminated all the outliers presented in the dataset for a total of 447 rows.



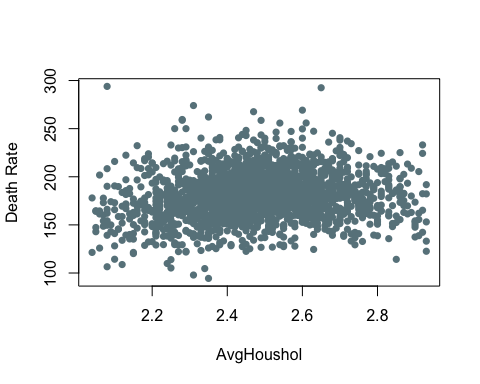
### Visualization of trends

In the next section we built scatterplots for each of the predictors against the Death Rate dependent variable. This task allowed us to discover any eventual linear trendline. Also we performed a R printout for the coefficient of variation.



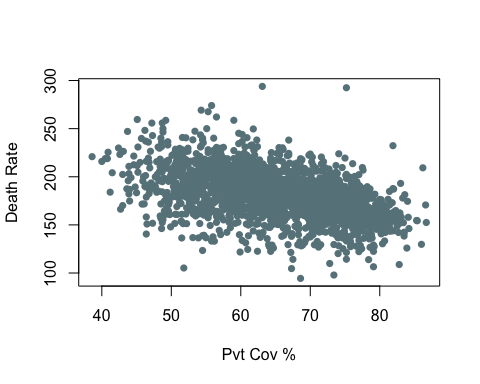
## [1] 0.4530695

We can see from the plot that there is a somewhat weak positive linear relationship between Poverty Percentage and Death Rate where as the Poverty Percentage increase so does Death Rate but not by a wide margin. The coefficient of correlation here is 0.4530695



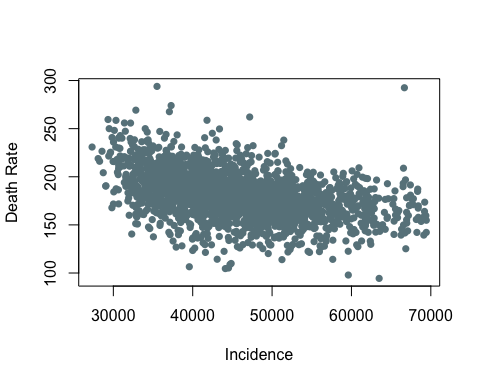
## [1] 0.155252

We can see from the plot that there is a somewhat weak positive linear relationship between Average Household Size and Death Rate where as the Average household Size increase so does Death Rate although very slightly. The coefficient of correlation here is in fact low at 0.155252.



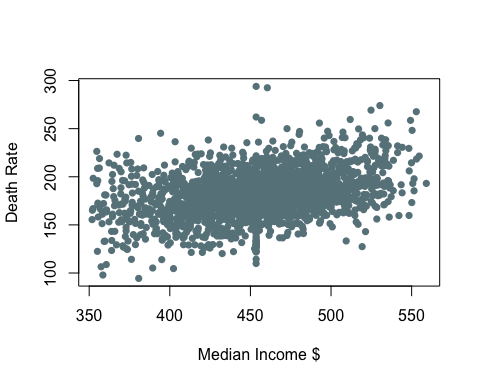
## [1] -0.4569882

We can see from the plot that there is a somewhat weak negative linear relationship between Private Coverage Percentage and Death Rate where as the Private Coverage Percentage increases the Death Rate actually decreases somewhat. The coefficient of correlation here is -0.4569882



## [1] -0.4461167

We can see from the plot that there is a somewhat weak negative linear relationship between Incidence Rate and Death Rate where as the Incidence Rate increases the Death Rate actually decreases somewhat. The coefficient of correlation here is -0.446116



## [1] 0.3519352

We can see from the plot that there is a farily weak positive linear relationship between Median Income and Death Rate where as the Median Income increases the Death Rate also increases somewhat. The coefficient of correlation here is 0.3519352

### 

### Dataframe split

There are 2154 rows in our dataframe. We have randomly selected 80% of rows from our dataframe for a total of 1723 data. Those data are now parts of a Train dataset we will use to perform regression analysis from now on. At the same time we used the remaining 431 data for testing. We created two .txt files with data formatted as table and uploaded them on GitHub. This RMarkdown reads data directly from Github repository.

For the next part we created a multiple linear regression model to predict cancer death rate using our predictors. It follows a R printout

##   
## Call:  
## lm(formula = Traincancer\_data$TARGET\_deathRate ~ Traincancer\_data$povertyPercent +   
## Traincancer\_data$AvgHouseholdSize + Traincancer\_data$PctPrivateCoverage +   
## Traincancer\_data$incidenceRate + Traincancer\_data$medIncome +   
## as.factor(Traincancer\_data$Region.CAT.), data = Traincancer\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -71.103 -10.872 -0.368 10.724 107.361   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 1.427e+02 1.191e+01 11.985  
## Traincancer\_data$povertyPercent -9.510e-02 1.822e-01 -0.522  
## Traincancer\_data$AvgHouseholdSize 9.557e+00 3.208e+00 2.979  
## Traincancer\_data$PctPrivateCoverage -7.079e-01 9.819e-02 -7.209  
## Traincancer\_data$incidenceRate 2.104e-01 1.152e-02 18.273  
## Traincancer\_data$medIncome -7.072e-04 1.204e-04 -5.872  
## as.factor(Traincancer\_data$Region.CAT.)Northeast -6.373e+00 1.747e+00 -3.649  
## as.factor(Traincancer\_data$Region.CAT.)South 2.120e+00 1.257e+00 1.686  
## as.factor(Traincancer\_data$Region.CAT.)West -1.575e+01 1.777e+00 -8.861  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## Traincancer\_data$povertyPercent 0.601789   
## Traincancer\_data$AvgHouseholdSize 0.002935 \*\*   
## Traincancer\_data$PctPrivateCoverage 8.40e-13 \*\*\*  
## Traincancer\_data$incidenceRate < 2e-16 \*\*\*  
## Traincancer\_data$medIncome 5.16e-09 \*\*\*  
## as.factor(Traincancer\_data$Region.CAT.)Northeast 0.000271 \*\*\*  
## as.factor(Traincancer\_data$Region.CAT.)South 0.091912 .   
## as.factor(Traincancer\_data$Region.CAT.)West < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.36 on 1714 degrees of freedom  
## Multiple R-squared: 0.4256, Adjusted R-squared: 0.4229   
## F-statistic: 158.8 on 8 and 1714 DF, p-value: < 2.2e-16

= Poverty Percent = -0.0951, when Poverty Percent is increased by 1 percent, the average Cancer Death Rate decreases by 0.0951 deaths per 1000 individuals per year, keeping other variables fixed.

= Avg Household Size = 9.557, when Average Househould Size is increased by 1 person, the average Cancer Death rate increases by 9.557 deaths per 1000 individuals per year.

= Percentage Private Coverage = -0.7079, when Percentage Private Coverage increases by 1 percent, the average Cancer Death Rate decreases by 0.7079 deaths per 1000 individuals per year.

= Incidence Rate = 0.2104, when Incidence Rate increases by 1, the average Cancer Death Rate increases by 0. 2104, deaths per 1000 individuals per year.

= Median Income = -0.0007072, when Median Income increase by 1 unit, the average Cancer Death Rate decreases by 0.0007072 deaths per 1000 individuals per year.

= Intercept = 142.7, the intercept value is not relevant in this model because there cannot be households of zero, as well a median income of 0$.

Testing Hypothesis for slopes

is not useful in the model

is useful in the model

As we look at the parameters from the printout that we used to predict the cancer death rate we can see by looking at the p-values of each parameter: Percentage of Private Coverage, Incidence Rate, Median Income and Average Household Size all have p-vaules that are significantly less than 5 percent. So we can say that the data provides evidence at a 5 percent significance level to say that Average Househol Size (p-value = 0.002935 < 0.05), Percentage Private Coverage (p-value = 8.40e-13 < 0.05), Incidence Rate (p-value = 2e-16 < 0.05), and Median income (p-value = 5.16e-09 < 0.05) are all useful variables in the model that affect Cancer Death Rate. However, Poverty Percent (p-value = 0.601789 > 0.05) is not useful in the model and thus does not affect Cancer Death Rate

The standard error of regression is 18.36 telling us that the average difference of the observed cancer death rate and the estimated Cancer Death Rate is 18.36

The coefficient of variation is 18.36\*100/mean(Traincancer\_data$TARGET\_deathRate) which is 10.12 percent, which tells us the estimated standard error of the regression model is 10.12 percent of the mean of the observed Cancer Death Rate. Since the CV is 10.12 percent which is slightly greater than 10 percent, the regression model would not be favorable although it is very close to 10 percent.

The multiple coefficient of determination is 0.4256, which tells us that about 42.56 percent of the sample variation in the Cancer Death Rate is explained by Poverty Percent, Average Household Size, Percetage Private Coverage, Incidence rate, and Median Income in the multiple linear regression model.