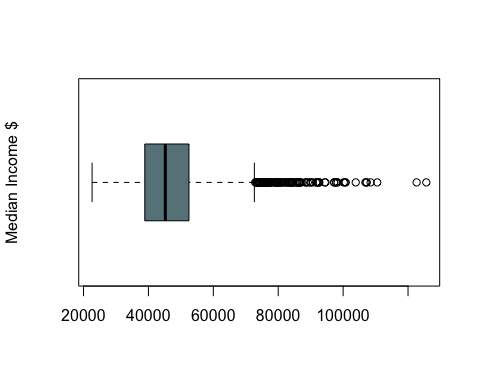
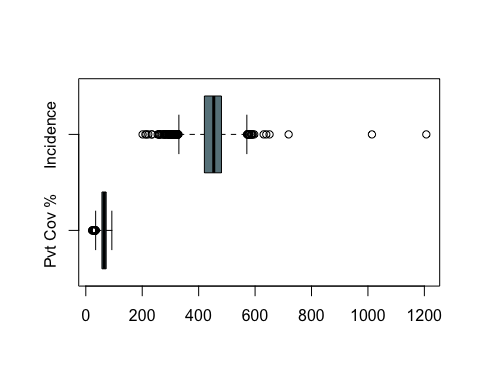
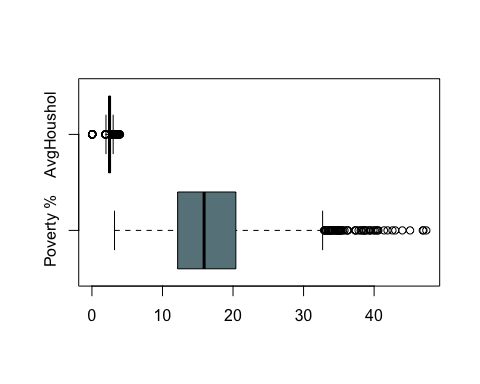
Group Project

field1, field2, field3

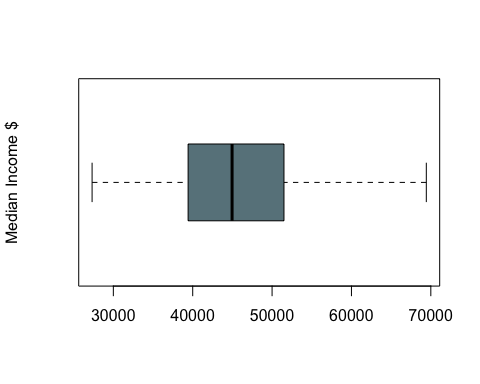
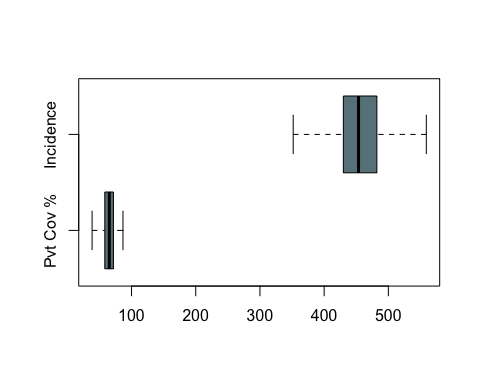
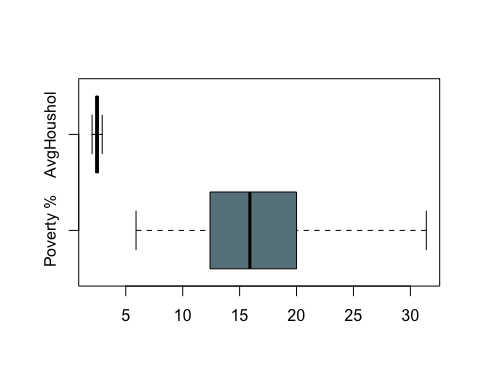
10/6/2021

# Project



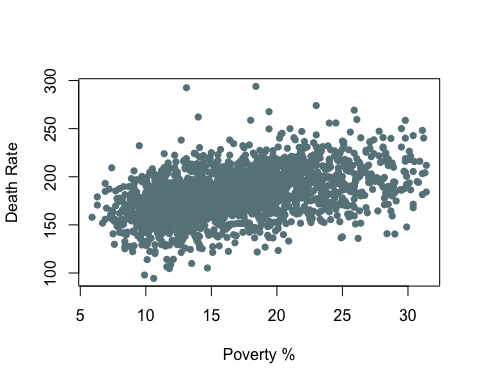
### Write something about it

We are going to remove outliers

 ### We removed all of the 447 outliers from dataset!

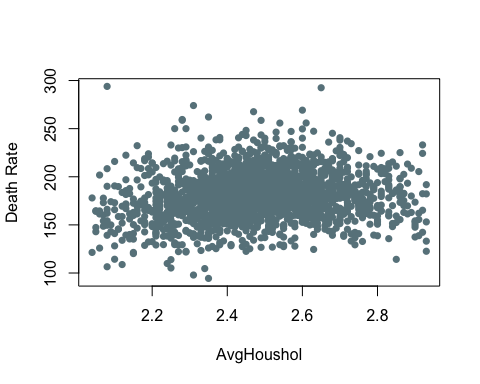
# Plots between predictors and death rate with their respective correlation coefficients

We need to write something underneath each of the plots



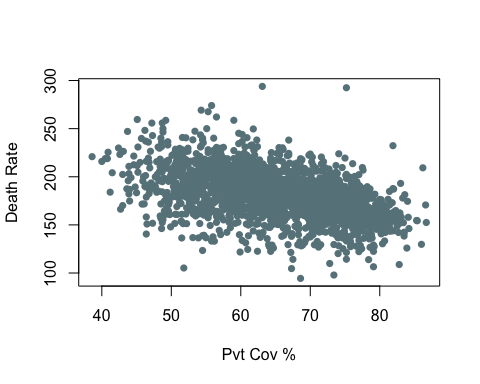
## [1] 0.4530695

### Write something here about this



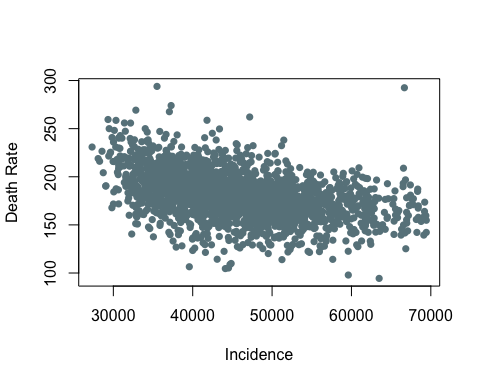
## [1] 0.155252

### Write something here about this



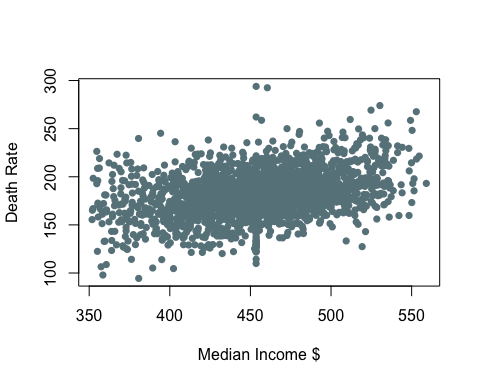
## [1] -0.4569882

### Write something here about this



## [1] -0.4461167

### Write something here about this



## [1] 0.3519352

### Write something here about this

### There are 2154 rows in our dataframe. We have randomly select 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. At the same time we used the remaining 431 data for testing.

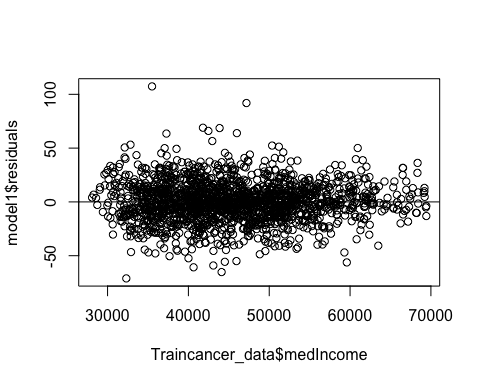
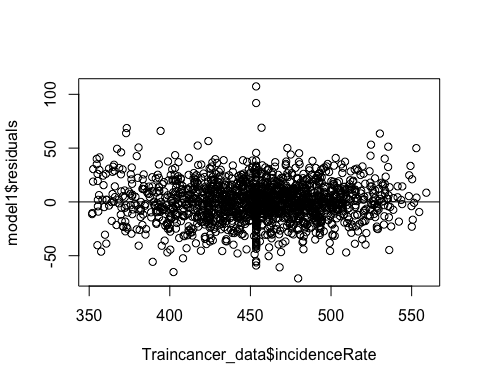
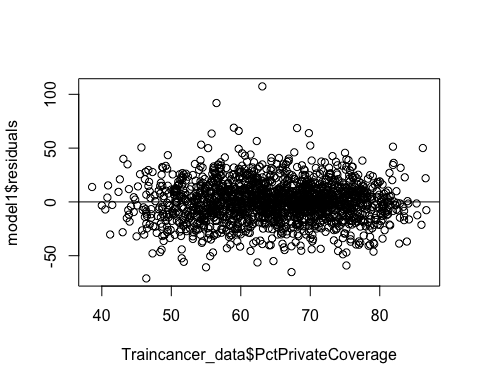
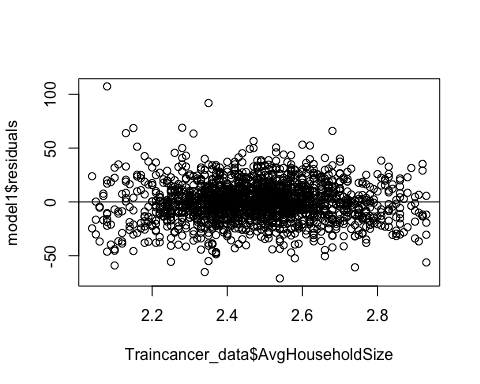
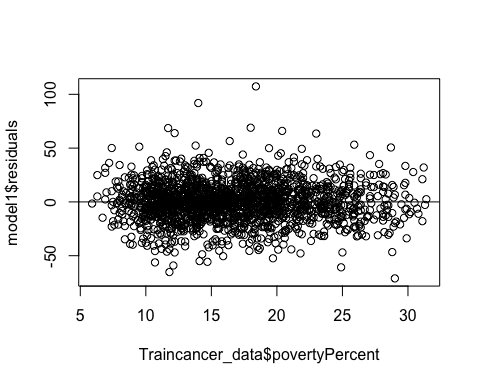
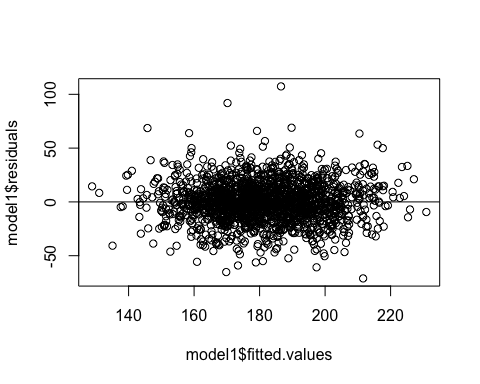
### Created two .txt files with data formatted as table. Uploaded on GitHub. This RMarkdown reads data directly from Github repository

## Now we are creating a multiple linear regression model to predict cancer death rate using our predictors

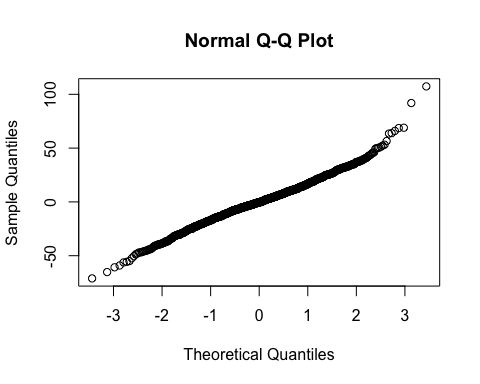
##   
## Call:  
## lm(formula = Traincancer\_data$TARGET\_deathRate ~ Traincancer\_data$povertyPercent +   
## Traincancer\_data$AvgHouseholdSize + Traincancer\_data$PctPrivateCoverage +   
## Traincancer\_data$incidenceRate + Traincancer\_data$medIncome +   
## as.factor(Traincancer\_data$Region.CAT.), data = Traincancer\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -71.103 -10.872 -0.368 10.724 107.361   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 1.427e+02 1.191e+01 11.985  
## Traincancer\_data$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

# Multiple Regression Model: Checking Assumptions

## Check Assumption 1 and 2: Check mean of error is zero, and check that the errors have the same variance given any combination of values of predictors, in doing so we plot residual plots:

 ### Assumption 1 conclusions: All the above residual plots are approximately symmetric around x-axis. The point look random and not patterns seen. So the assumption that mean of error is zero is satified. ### Assumption 2 conclusion: All residuals plot have an approximately horizontal band around x axis. therefore we can say that the error variance is constant for all different value combinations of predictors.

## Check Assumption 3: For any given set of predictor variables, the error is normally distributed, thus response variable y is normally distributed as well. We plot the qqnorm graph.

 ### Assumption 3 conclusions: The qqplot of residuals looks approximately linear. Therefore the residuals of the errors are normally distributed.

## Check Assumption 4: The observations of the response variable are independent from one another. Observations are not correlated. We use either the DW test value or the p-value.

## Caricamento del pacchetto richiesto: lmtest

## Caricamento del pacchetto richiesto: zoo

##   
## Caricamento pacchetto: 'zoo'

## I seguenti oggetti sono mascherati da 'package:base':  
##   
## as.Date, as.Date.numeric

##   
## Durbin-Watson test  
##   
## data: model1  
## DW = 2.0029, p-value = 0.9512  
## alternative hypothesis: true autocorrelation is not 0

### Assumption 4 conclusions: The p-value is greater than 0.05 (0.9512 > 0.05) so we fail to reject H0: “Independent errors”. Therefore we will take the assumption as satisfied.

### We still don’t know how to check the 5th assumption: None of the predictors are constant and there are no exact linear relationships among predictors.

### Assuming all the multiple linear regression assumptions are satisfied. We can perform the individual t-test for each individual predictor

##   
## 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

### In the above output we can see that the individual t-values of, ‘AvgHouseholdSize’, ‘PctPrivateCoverage’, ‘incidenceRate’, and ‘medIncome’ are less than 0.05. So they are useful predictor variables for colGPA

### However, ‘povertyPercent’ predictor has a value of 0.6, greater than 0.05 so is not a useful predictor variable to predict the death rate.

###Lets create a new model without ACT variable and check which model is better, the one with ‘povertyPercent’ variable (model1), or the one without ACT variable (model2). We will perform an AIC (Alike Information Criterion) test, where the smaller AIC is preferred.

## [1] 14929.99

## [1] 15041.83

### The model 2 has a lower AIC score, therefore we are going to eliminate the ‘povertyPercent’ predictor variables from our model.

#Check prediction accuracy for the improved model.

###We need first to predict the response variable value using a new dataset, which is the TestCancerData.txt dataset that is in the GitHub repository. Then we will use the MAPE, Mean Absolute Percentage Error (MAPE)

## Warning: 'newdata' ha 431 righe ma la variabile trovata ha 1723 righe

# Checking MultiCollinearity

#### We are going to follow 4 ways to check multicollinearity

## 1. Using correlation between predictor variables

## Caricamento del pacchetto richiesto: knitr

Correlation Between Predictors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AvgHouseholdSize | PctPrivateCoverage | incidenceRate | medIncome |
| AvgHouseholdSize | 1.000 | -0.285 | -0.010 | 0.049 |
| PctPrivateCoverage | -0.285 | 1.000 | 0.061 | 0.770 |
| incidenceRate | -0.010 | 0.061 | 1.000 | 0.044 |
| medIncome | 0.049 | 0.770 | 0.044 | 1.000 |

### We see that (as expected) the ‘medIncome’ and ‘PctPrivateCoverage’ have a pretty high positive correlation with r values of 0.77 close to 1. So multicollinearity exists in the dataset and we need to drop one of these predictors. We can do that by creating separate models with one of those variables at the time and compering their respective AIC values, where smaller AIC is preferred.

## [1] 15145.84

## [1] 15067.53

### AIC in testModel2, the one without the’PctPrivateCoverage’ predictor variable, is smaller than AIC of testModel1, so we drop the ‘PctPrivateCoverage’ variable in the model

## 2. Opposite signs in parameter estimates than what is expected