STAT1006

Regression and Nonparametric Inference

Semester 2, 2023

Final Project Report

Linear Regression Model of Cancer Mortality in a Country

Supawit Praditkul, 21657005

Bachelor of Science (Data Science)

Declaration

The work presented in this report is my own work and all references are duly acknowledged.

This work has not been submitted, in whole or in part, in respect of any academic award at Curtin University or elsewhere.

Supawit Praditkul

19/10/2023

Contents

1.Introduction

According to Australian bureau of statistics (2016), cancer is a group of abnormal cells that multiply uncontrol in a body, which leading them to steal the resource from other part of the body. If the cancer is not treated in an early stage, it could lead to loss of patient life.

By understanding what variables is affect the cancer mortality with linear regression model, not only prediction can be made with the model, but each county can look at the model at try to reduce the cancer mortality by increase or decrease variables that have a strong relation according to the model.

The data used in this work is the cancer mortality base on United State County with the total of 2000 data point with total of 34 Variables including mean mortality per capita (100,000) In which we will create the model to predict this variable.

This work goal is to create two linear regression model that best represent the mean mortality per capita, each for single linear regression and multiple linear regression from the data.

2.Exploratory Data Analysis

After exploring the mean mortality per capital, all the data point looks fine. They are normally distributed and some from lots of different US county. There might be some potential outlier but it not abnormal or impossible for the data to have this value.

The variable I will not make the model for is the Geography variable which represent the US county which the data come from, the average number of reported cases of cancer diagnosed annually, mean number of reported mortalities due to cancer per year, and incidence rate since the last there is dependent to the variable, we want to explore.

There are some data point that don't have all the value for every variable. The Percent of county residents ages 18-24 highest education attained some college (PctSomeCol18_24) specifically have only about 500 data point which is very low data point. Decision to not take this variable into consideration in making the model for both single linear regression and multiple variable linear regression is made. The others variable that don't have full data point are percent of county residents ages 16 and over employed (PctEmployed16_Over), percent of county residents with private health coverage alone (PctPrivateCoverageAlone), and percent of county residents with employee-provided private health coverage (PctEmpPrivCoverage). In these three viable, decision to continue make the model with the rest of the data in single linear

2.Exploratory Data Analysis (cont.)

regression is make and to work on multiple linear regression all the data point with no variable in at least one of those three needed to be cut off.

Next, is the outlier that have been cut before making the model. There lots of outlier in the data but most of them is kept due to they not effect to model that much, or it's a valid data point that could possibly happen but there are data in two variables that have been cut due to the data is not possible to happen, this might be due to some error in collecting the data or when enter to the data or how they store the data, either way contract with the data owner is not possible for this project so the decision to cut those data and continue make the model with the rest of the data in single linear regression is make and to work on multiple linear regression all the data point with outlier in at least one of them needed to be cut off.. The two variable that have those data are Mean household size of county (AvgHouseholdSize) and Median age of county residents (MedianAge) and follow will be how the outlier have been treat.

For the variable Mean household size of county, the definition of household according to Australian bureau of statistics (2016) quote "A household is defined as one or more persons, at least one of whom is at least 15 years of age, usually resident in the same private dwelling.". with this definition the average size of household cannot be less than one.

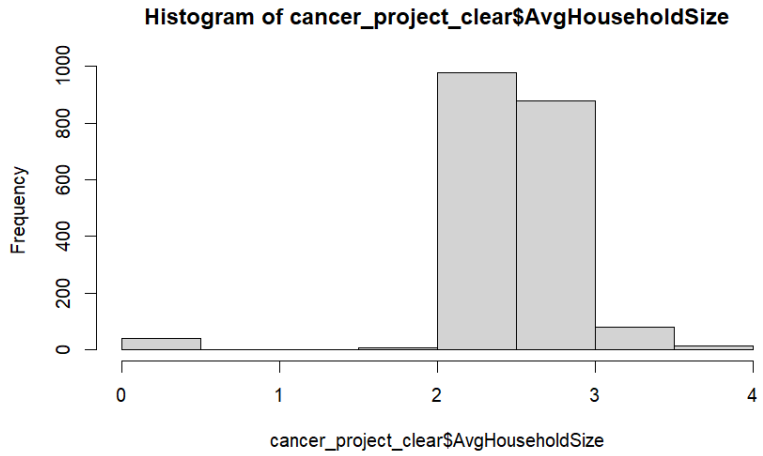

Figure 1: Histogram of Mean household size of county

From this information the decision to cut the data point of average house with the value of less than 1 is made.

For the variable Median age of county residents there are some data point with the value more than 200 years

2.Exploratory Data Analysis (cont.)

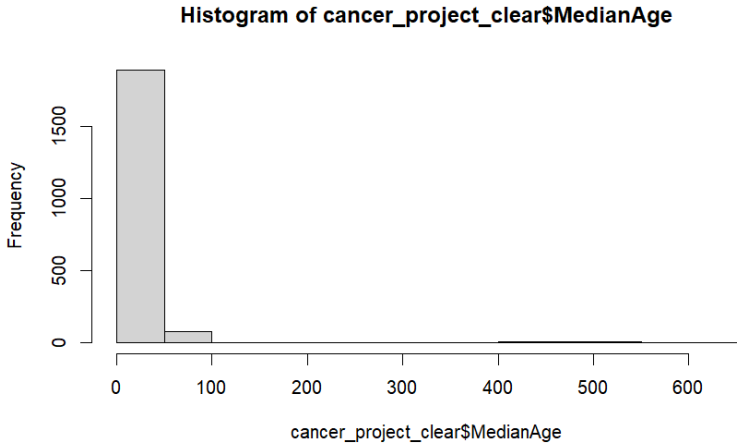

Figure 2: Histogram of Median age of county residents

Such data point should not possible so the decision of cutting the data point with more than 200 years have been made.

Next after looking at plot of mean mortality per capita against others variable and looking at their error have constant variance or that the value is normally distribute if not the data transformation will be performed. There are 14 variables that have been transform by this process using one of these transformations: natural logarithm, square root, 4^{th} root, 5^{th} root, and 4^{th} power.

The data in variables that undergo natural logarithm transformation are followed: median income per county (medIncome), population of county (popEst2015), mean household size of county (AvgHouseholdSize), and percent of county residents ages 25 and over highest education attained bachelor's degree (PctBachDeg25_Over).The data in variables that undergo Square root transformation are followed: percent of populace in poverty (povertyPercent), percent of county residents ages 18-24 highest education attained less than high school (PctNoHS18_24), and number of live births relative to number of women in county (BirthRate).The data in variables that undergo 4^{th} root transformation are followed: per capita number of cancer-related clinical trials per county (studyPerCap), percent of county residents ages 18-24 highest education attained: bachelor's degree (PctBachDeg18_24), and percent of county residents who identify as Black (PctBlack). The data in variables that undergo 5^{th} root transformation are followed: percent of county residents who identify as Asian (PctAsian), and percent of county residents who identify in a category which is not White, Black, or Asian (PctOtherRace). The data in variable that under go 4^{th} power transformation is percent of county residents who identify as White (PctWhite).

3.Simple Linear Regression: Analysis and Results

The following procedure is made in order to find the best represent single linear regression model for mean mortality per capita (100,000) represent as mortality

1. Make the plot of mortality against all other variable
2. Check with F-statistic, and ANOVA to confirm that the relationship of the model is valid, if not drop the model
3. Do the plot on the rest of the variables following the step 1-3
4. Store all the models' R-square and mean square error (MSE) that have been confidence test that they have a relationship and found the model that have best represent the interest variance base on The higher the R-square get, the better fit the model to the data point and, The less the MSE the less the better the model represent the interest variance.

The best fit model is a model between mean mortality per capita (mortality) and natural logarithm of percent of county residents ages 25 and over highest education attained: bachelor's degree (PctBachDeg25_Over). The model can be written as follow:

$$Mortality = 265.424 - 34.472 * log\ (PctBachDeg25\_Over)$$

The model tells that for every value of natural log of percent of county residents ages 25 and over highest education attained: bachelor's degree the mean mortality per capita is reduce by 34.472. Here the graph plot with the model above and the data that it tries to represent.

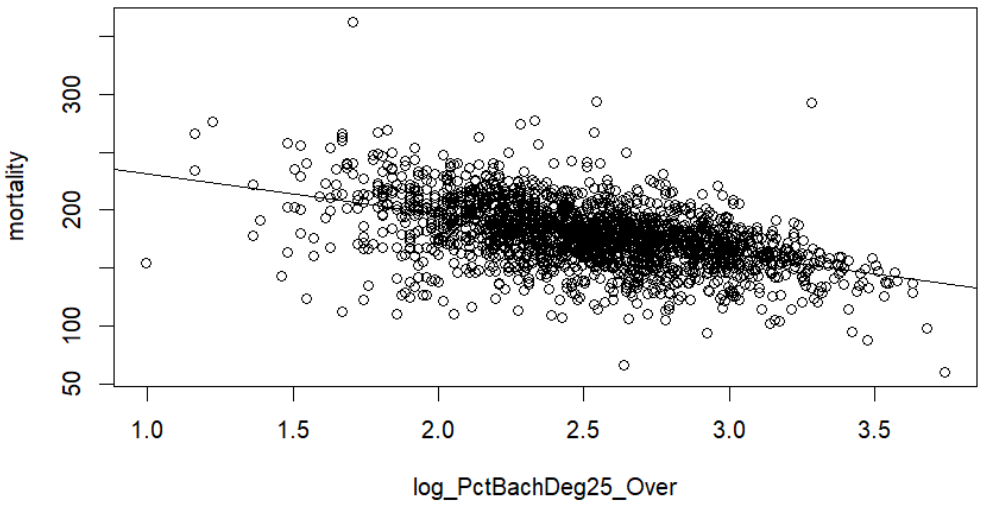

Figure 3: Plot of the best fit model with the data its try to fit

3.Simple Linear Regression: Analysis and Results (cont.)

After testing with F-test and ANOVA to confirm a relationship both return with very low P-value confirm that there is a relationship between these two variables.

Next is perform diagnostic for the model

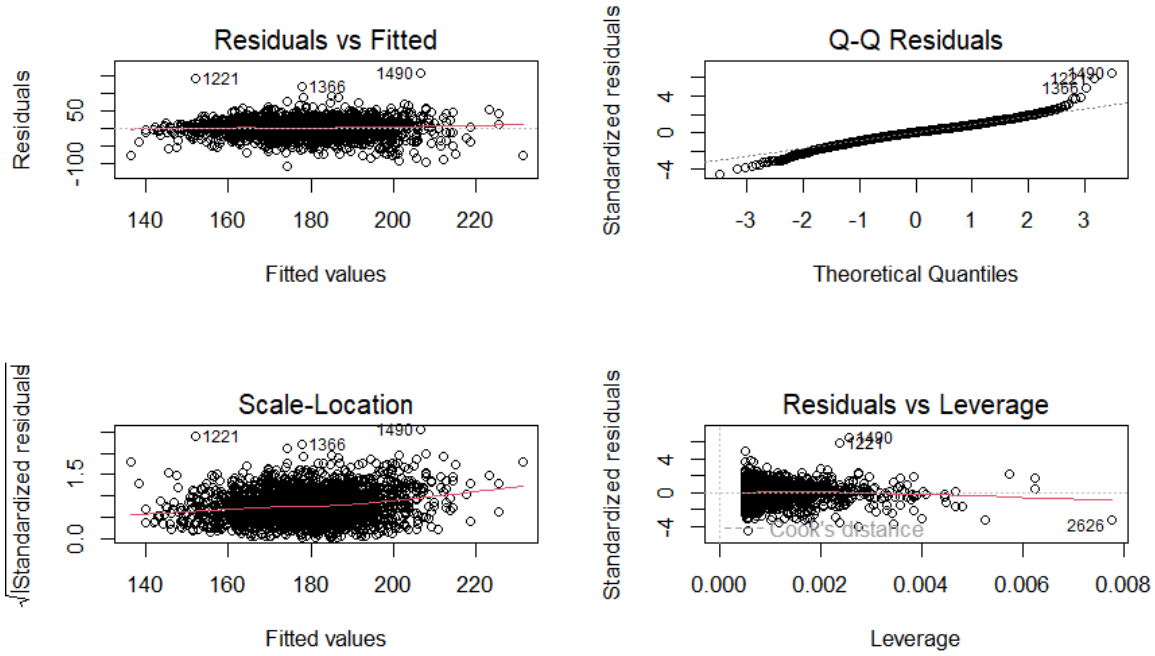

Figure 4-7: diagnostic plots of the best SLR model

From the diagnostic plot, the top left plot shows the residuals against fitted values to confirm that the data is random and have constant variance. The plot in bottom left father confirm that have constant variance with almost flat line in the plot. The top right plot is a normal Q–Q plot. In this plot, most of the observations lie around the straight line, confirming normality. The last plot of standardized residual against leverage show the outlier in this plot there are total to three outliers in this model but decision to not cut those outliers is made due to the data point is possible to be true and the researcher want the model to represent these data as well.

This model has the highest R-square of 0.2469 and Lowest mean square error (MSE) of 575.38 making it the best model represent the interest variable the second and third best model are the model with natural log of median income per county and Median income per capita binned by decile both having R-square around 0.18 and MSE around 620 both significant differ from the best model.

4.Multiple Linear Regression: Analysis and Results

The best linear regression model discovered is below. From 28 predict variables and 1,482 after going through data cleaning and transformation

$$
\begin{aligned}
mortality = {} & 248.67 - 17.029\log(PctBachDeg25_{Over}) \\
& - 1.74PctMarriedHouseholds + 0.73PctHS25_{Over} \\
& + 8.02\sqrt[4]{PctBlack} - 12.18\sqrt[5]{PctOtherRace} \\
& + 1.77\log(popEst2015) - 6.57\sqrt{Birthrate} + 0.17PctHS18_{24} \\
& - 0.94MedianAgeFemale + 1.6PercentMarried \\
& - 4.66\sqrt[4]{PctBachDeg18_{24}} - 2.23\sqrt{PctNoHS18_{24}} \\
& + 0.85PctEmpPrivCoverage - 0.44PctEmployed16_{Over} \\
& - 0.8PctPrivateCoverageAlone
\end{aligned}
$$

The model tells that for every value of the variance after transformation multiply by its own parameter will affect mean mortality per capita by that amount. In total 15 variance have a significant effect to mean mortality per capita.

This model is achieved by using forward selection. After that, using partial F-test for comparing models to remove some of the variable that when remove didn't show significant different between before and after remove.

After the model is construct, it undergoes diagnostic checking.

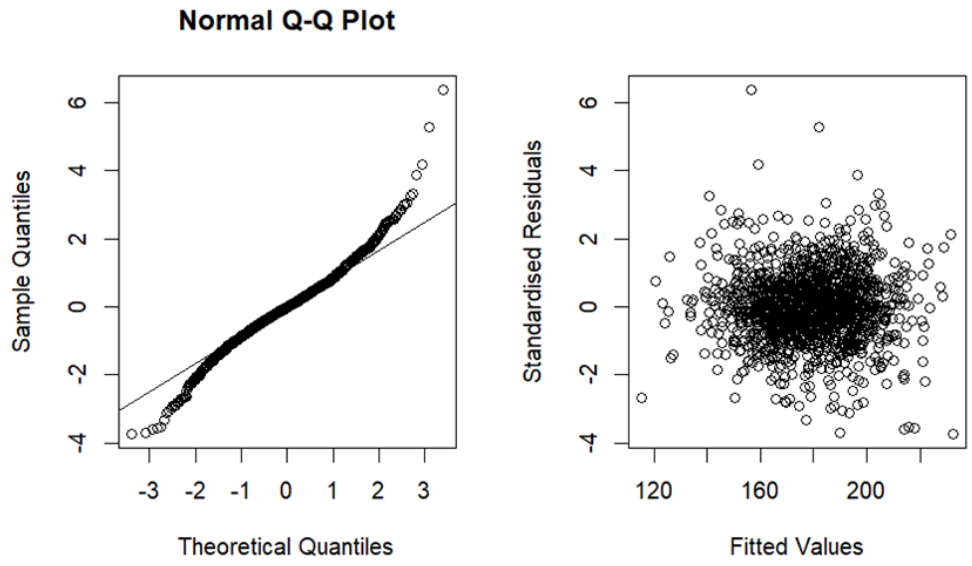

Figure 8-9: Q-Q plot of the standard residual and plot of standard residual against fitted value

From normal Q-Q plot, most of the observations lie around the straight line, confirming normality and from the plot on the right confirm the residual are random.

4.Multiple Linear Regression: Analysis and Results(cont.)

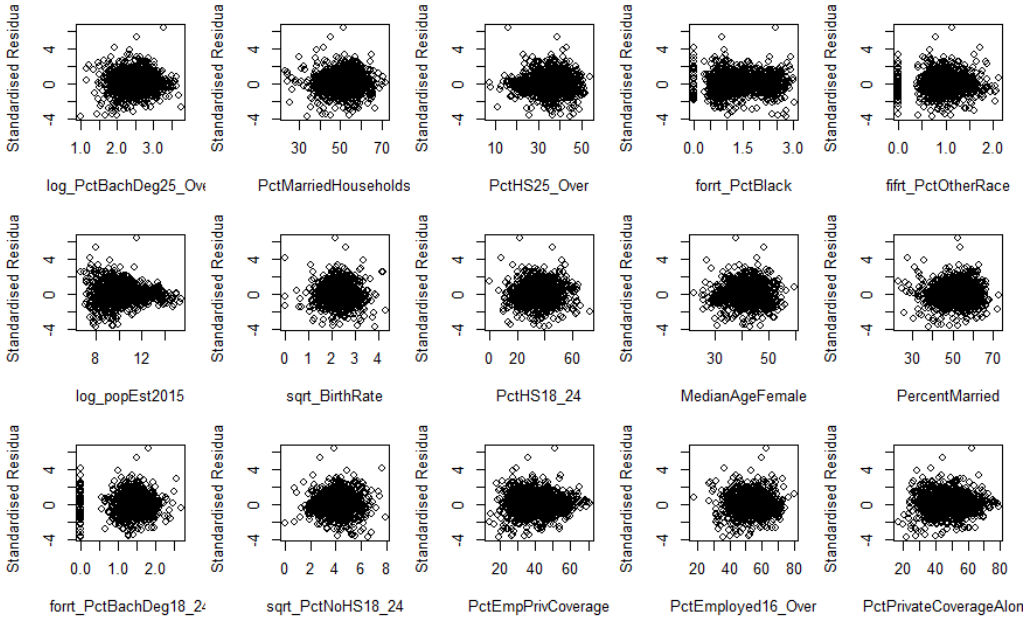

Figure 10-25: Plots of standardized residuals against each of explanatory variables

These plots further confirm the plots before, that the data is random and have constant variance.

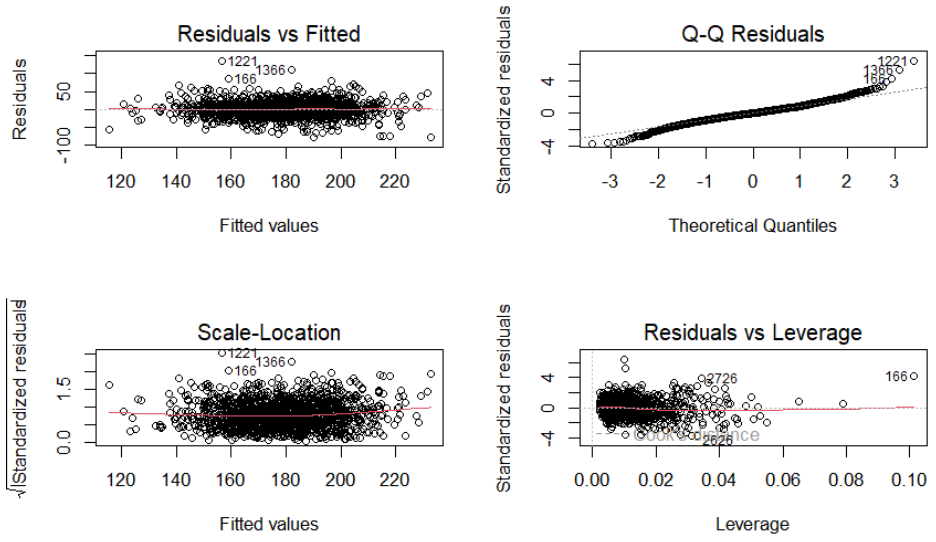

Figure 26-30: diagnostic plots of the best SLR model

The last plot of standardized residual against leverage show the outlier in this plot a total of 1 outlier with high leverage but decision to not cut those outliers is made due to the data point is possible to be true and the researcher want the model to represent these data as well.

5. Conclusion

Two model have been conducted in order to predict the mean mortality rate per capita. One is the Simple linear regression model, the other is Multiple linear regression model.

In simple linear regression model, a variable that was found to have influence on mean mortality rate per capita the most is percent of county residents ages 25 and over highest education attained: bachelor's degree after going to natural log transformation. Both variables have negative relationship meaning that if the percent of county residents ages 25 and over highest education attained: bachelor's degree increase, the mean mortality rate per capita will decrease.

In multiple linear regression model, there are 15 variables that represent mean mortality rate per capita. The variable that have the most influent is percent of county residents ages 25 and over highest education attained: bachelor's degree after going to natural log transformation. Both variables have negative relationship meaning that if the percent of county residents ages 25 and over highest education attained: bachelor's degree increase, the mean mortality rate per capita will decrease.

This could be assuming that if resident has at least a bachelor's degree could help them diagnose the cancer at the earlier stage and able to get the treatment in time leading to low mortality rate in overall county. Further research has to been conduct to confirm this theory.

In this research, the things that want to improve is the lost data that un able to obtain, specifically the data in percent of county residents ages 18-24 highest education attained some college (PctSomeCol18_24) that have very low data point and unable to make reasonable model between the interest variable. If the contract to the data gatherer can be done, the data point might be able to recover hand might conduct a model that better represent the mean mortality rate per capita than models that discover in this research.

References

Australian bureau of statistics. (2016, August 23). *Chapter - household*. Www.abs.gov.au. https://www.abs.gov.au/ausstats/abs@.nsf/lookup/2901.0chapter34902016

Cancer Council. (2019). What is cancer? Cancer Council. https://www.cancer.org.au/cancer-information/what-is-cancer

Appendix

R code from R-markdown

Final Project

Supawit Praditkul

2023-10-13

```
load("cancer.RData")
```

```
head(cancer_project)
```

```
str(cancer_project)
```

```
#remove the bininc
cancer_project_remove_bininc <- cancer_project[-c(9)]
#show all the decile
sort(unique(cancer_project$binnedInc))
#create new value for decile
new_bininc = c()
for (x in cancer_project$binnedInc) {
  if ((x == "[22640, 34218.1]")){
    new_bininc <- append(new_bininc,values = 1)
  }
    if ((x == "(34218.1, 37413.8]")){
    new_bininc <- append(new_bininc,values = 2)
    }
    if ((x == "(37413.8, 40362.7]")){
    new_bininc <- append(new_bininc,values = 3)
    }
    if ((x == "(40362.7, 42724.4]")){
    new_bininc <- append(new_bininc,values = 4)
    }
    if ((x == "(42724.4, 45201]")){
    new_bininc <- append(new_bininc,values = 5)
    }
    if ((x == "(45201, 48021.6]")){
    new_bininc <- append(new_bininc,values = 6)
    }
    if ((x == "(48021.6, 51046.4]")){
    new_bininc <- append(new_bininc,values = 7)
    }
    if ((x == "(51046.4, 54545.6]")){
    new_bininc <- append(new_bininc,values = 8)
    }
    if ((x == "(54545.6, 61494.5]")){
    new_bininc <- append(new_bininc,values = 9)
    }
    if ((x == "(61494.5, 125635]")){
```

```
    new_bininc <- append(new_bininc,values = 10)
    }
}
cancer_project_clear <- cbind(cancer_project_remove_bininc,
                              BinnedInc = new_bininc)
```

```
boxplot(cancer_project_clear$mortality)
hist(cancer_project_clear$mortality)
head(sort(cancer_project_clear$mortality,decreasing = T))
summary(cancer_project_clear$mortality)
#the data look normally distribute
#might be outlier in the data
#US County Florida, Presidio County Texas and Pitkin County, Colorado
```

```
mortal_medin_lm <- lm(mortality~medIncome, data = cancer_project_clear)
summary(mortal_medin_lm)
plot(mortality~medIncome, data = cancer_project_clear)
abline(mortal_medin_lm)
plot(mortal_medin_lm)
hist(cancer_project_clear$medIncome)

#the data is skew
```

```
#try transform the data to log of incidence rate
log_medIncome <- log(cancer_project_clear$medIncome)

mortal_logmedIncome_lm <- lm(mortality~log_medIncome,
                             data = cancer_project_clear)
summary(mortal_logmedIncome_lm)
plot(mortality~log_medIncome, data = cancer_project_clear)
abline(mortal_logmedIncome_lm)
plot(mortal_logmedIncome_lm)
hist(log_medIncome)

#find outlier
Leverage <- hatvalues(mortal_logmedIncome_lm)
plot(Leverage ~ log_medIncome, data = cancer_project_clear,
     xlab = "log_medIncome", ylab = "Leverage")
abline(h = 4/(length(log_medIncome)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_logmedIncome_lm)
plot(Cooks.Dist ~ log_medIncome, data = cancer_project_clear,
     xlab = "log_medIncome", ylab = "Cook's distance")
abline(h = 4/(length(log_medIncome)-2), col = "red")
#their is outliers but it doesn't have big gap from the main group
#decide not to cut the outliers

#t test for  relationship
#pval is low
#their might be no relationship

#anova test
#the data look suit for anova
```

```r
anova(mortal_logmedIncome_lm)
#their might be a relationship
```

```r
mortal_pop_lm <- lm(mortality~popEst2015, data = cancer_project_clear)
summary(mortal_pop_lm)
plot(mortality~popEst2015, data = cancer_project_clear)
abline(mortal_pop_lm)
plot(mortal_pop_lm)
hist(cancer_project_clear$popEst2015)
#the data is not normally distribute

#transformation is needed
```

```r
#try transform the data to log of incidence rate
log_popEst2015 <- log(cancer_project_clear$popEst2015)

mortal_logpopEst2015_lm <- lm(mortality~log_popEst2015,
                              data = cancer_project_clear)
summary(mortal_logpopEst2015_lm)
plot(mortality~log_popEst2015, data = cancer_project_clear)
abline(mortal_logpopEst2015_lm)
plot(mortal_logpopEst2015_lm)
hist(log_popEst2015)

#find outlier
Leverage <- hatvalues(mortal_logpopEst2015_lm)
plot(Leverage ~ log_popEst2015, data = cancer_project_clear,
     xlab = "log_popEst2015", ylab = "Leverage")
abline(h = 4/(length(log_popEst2015)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_logpopEst2015_lm)
plot(Cooks.Dist ~ log_popEst2015, data = cancer_project_clear,
     xlab = "log_popEst2015", ylab = "Cook's distance")
abline(h = 4/(length(log_popEst2015)-2), col = "red")
#their are outliers but thier isn't that much gap from the main group
#decide not to cut the outliers

#t test for negative relationship
#pval is low
#there is  relationship

#anova test for relation
#the data look suitable for anova
anova(mortal_logpopEst2015_lm)
#there is relationship
```

```r
mortal_poor_lm <- lm(mortality~povertyPercent, data = cancer_project_clear)
summary(mortal_poor_lm)
plot(mortality~povertyPercent, data = cancer_project_clear)
abline(mortal_poor_lm)
plot(mortal_poor_lm)
hist(cancer_project_clear$povertyPercent)
# the data is skew consider transfrom the data
```

```r
#try transform the data to sqrt of incidence rate
sqrt_povertyPercent <- sqrt(cancer_project_clear$povertyPercent)

mortal_sqrtpovertyPercent_lm <- lm(mortality~sqrt_povertyPercent,
                                   data = cancer_project_clear)
summary(mortal_sqrtpovertyPercent_lm)
plot(mortality~sqrt_povertyPercent, data = cancer_project_clear)
abline(mortal_sqrtpovertyPercent_lm)
plot(mortal_sqrtpovertyPercent_lm)
hist(sqrt_povertyPercent)

#find outlier
Leverage <- hatvalues(mortal_sqrtpovertyPercent_lm)
plot(Leverage ~ sqrt_povertyPercent, data = cancer_project_clear,
     xlab = "sqrt_povertyPercent", ylab = "Leverage")
abline(h = 4/(length(sqrt_povertyPercent)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_sqrtpovertyPercent_lm)
plot(Cooks.Dist ~ sqrt_povertyPercent, data = cancer_project_clear,
     xlab = "sqrt_povertyPercent", ylab = "Cook's distance")
abline(h = 4/(length(sqrt_povertyPercent)-2), col = "red")

#their is an outlier but the gap isn't that big and data look reasonable
#decide not to cut the outlier

#t test for negative relationship
#pval is low
#there is  relationship

#anova test for relation
#the data look suitable for anova
anova(mortal_sqrtpovertyPercent_lm)
#there is relationship
```

```r
mortal_study_lm <- lm(mortality~studyPerCap, data = cancer_project_clear)
summary(mortal_study_lm)
plot(mortality~studyPerCap, data = cancer_project_clear)
abline(mortal_study_lm)
plot(mortal_study_lm)
hist(cancer_project_clear$studyPerCap)
#the data is not normally distributed
```

```r
#try transform the data to 4th root of incidence rate
forrt_studyPerCap <- (cancer_project_clear$studyPerCap)^0.25

mortal_forrtstudyPerCap_lm <- lm(mortality~forrt_studyPerCap,
                                 data = cancer_project_clear)
summary(mortal_forrtstudyPerCap_lm)
plot(mortality~forrt_studyPerCap, data = cancer_project_clear)
abline(mortal_forrtstudyPerCap_lm)
plot(mortal_forrtstudyPerCap_lm)
hist(forrt_studyPerCap)
```

```r
#find outlier
Leverage <- hatvalues(mortal_forrtstudyPerCap_lm)
plot(Leverage ~ forrt_studyPerCap, data = cancer_project_clear,
     xlab = "forrt_studyPerCap", ylab = "Leverage")
abline(h = 4/(length(forrt_studyPerCap)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_forrtstudyPerCap_lm)
plot(Cooks.Dist ~ forrt_studyPerCap, data = cancer_project_clear,
     xlab = "forrt_studyPerCap", ylab = "Cook's distance")
abline(h = 4/(length(forrt_studyPerCap)-2), col = "red")
#their are some outlier but the data look valid
#decide not to cut the outlier

#t test for negative relationship
#pval is low
#there is  relationship

#anova test for relation
#the data not suitable for anova
```

```r
mortal_medage_lm <- lm(mortality~MedianAge, data = cancer_project_clear)
summary(mortal_medage_lm)
plot(mortality~MedianAge, data = cancer_project_clear)
abline(mortal_medage_lm)
plot(mortal_medage_lm)
hist(cancer_project_clear$MedianAge)
#there no way med age is above 300 years
#they are also bad leverage point
```

```r
#cleaning data
mortality_medage2 <- cancer_project_clear$mortality[
  cancer_project_clear$MedianAge<200]
MedianAge2 <-  cancer_project_clear$MedianAge[
  cancer_project_clear$MedianAge<200]

#construck new model
mortal_medage_lm2 <- lm(mortality_medage2~MedianAge2)
summary(mortal_medage_lm2)
plot(mortality_medage2~MedianAge2)
abline(mortal_medage_lm2)
plot(mortal_medage_lm2)

#find outlier
Leverage <- hatvalues(mortal_medage_lm2)
plot(Leverage ~ MedianAge2, data = cancer_project_clear,
     xlab = "MedianAge2", ylab = "Leverage")
abline(h = 4/(length(MedianAge2)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_medage_lm2)
plot(Cooks.Dist ~ MedianAge2, data = cancer_project_clear,
     xlab = "MedianAge2", ylab = "Cook's distance")
abline(h = 4/(length(MedianAge2)-2), col = "red")
#their some outlier but not enough gap to the main data group
```

```r
#decide not to cut the outlier

#t test for  relationship
#pval = 0.371
#their might be no relationship

#anova test
hist(MedianAge2)
#data suit for anova
anova(mortal_medage_lm2)
#their might be  no relation ship
```

```r
mortal_medageM_lm <- lm(mortality~MedianAgeMale, data = cancer_project_clear)
summary(mortal_medageM_lm)
plot(mortality~MedianAgeMale, data = cancer_project_clear)
abline(mortal_medageM_lm)
plot(mortal_medageM_lm)
hist(cancer_project_clear$MedianAgeMale)

#find outlier
Leverage <- hatvalues(mortal_medageM_lm)
plot(Leverage ~ MedianAgeMale, data = cancer_project_clear,
     xlab = "MedianAgeMale", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$MedianAgeMale)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_medageM_lm)
plot(Cooks.Dist ~ MedianAgeMale, data = cancer_project_clear,
     xlab = "MedianAgeMale", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$MedianAgeMale)-2), col = "red")
#there are some outliers but not enough gap from the main data group
#decide not to cut the outliers

#t test for  relationship
#pval = 0.175
#their might be no relationship

#anova test
#look suitable for anova test
anova(mortal_medageM_lm)
#their might be no relationship
```

```r
mortal_medageF_lm <- lm(mortality~MedianAgeFemale, data = cancer_project_clear)
summary(mortal_medageF_lm)
plot(mortality~MedianAgeFemale, data = cancer_project_clear)
abline(mortal_medageF_lm)
plot(mortal_medageF_lm)
hist(cancer_project_clear$MedianAgeFemale)

#find outlier
Leverage <- hatvalues(mortal_medageF_lm)
plot(Leverage ~ MedianAgeFemale, data = cancer_project_clear,
     xlab = "MedianAgeFemale", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$MedianAgeFemale)-2), col = "red")
```

```r
Cooks.Dist <- cooks.distance(mortal_medageF_lm)
plot(Cooks.Dist ~ MedianAgeFemale, data = cancer_project_clear,
     xlab = "MedianAgeFemale", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$MedianAgeFemale)-2), col = "red")
#there are some outliers but not enough gap from the main data group
#decide not to cut the outliers

#t test for  relationship
#pval = 0.116
#their might be no relationship

#anova
#data look suitable for anova
anova(mortal_medageF_lm)
#their might be no relationship
```

```r
mortal_housesize_lm <- lm(mortality~AvgHouseholdSize,
                          data = cancer_project_clear)
summary(mortal_housesize_lm)
plot(mortality~AvgHouseholdSize, data = cancer_project_clear)
abline(mortal_housesize_lm)
plot(mortal_housesize_lm)
hist(cancer_project_clear$AvgHouseholdSize)
#there noway household size is zero
```

```r
#cleaning data
mortality_housesize2 <- cancer_project_clear$mortality[
  cancer_project_clear$AvgHouseholdSize>1]
AvgHouseholdSize2 <-  cancer_project_clear$AvgHouseholdSize[
  cancer_project_clear$AvgHouseholdSize>1]

#new model
mortal_housesize_lm2 <- lm(mortality_housesize2~AvgHouseholdSize2)
summary(mortal_housesize_lm2)
plot(mortality_housesize2~AvgHouseholdSize2)
abline(mortal_housesize_lm2)
plot(mortal_housesize_lm2)
hist(AvgHouseholdSize2)

#the data look skew
#transform the data
```

```r
#try transform the data to log of AvgHouseholdSize2
log_AvgHouseholdSize2 <- log(AvgHouseholdSize2)

mortal_loghousesize_lm <- lm(mortality_housesize2~log_AvgHouseholdSize2)
summary(mortal_loghousesize_lm)
plot(mortality_housesize2~log_AvgHouseholdSize2)
abline(mortal_loghousesize_lm)
plot(mortal_loghousesize_lm)
hist(log_AvgHouseholdSize2)
```

```r
#find outlier
Leverage <- hatvalues(mortal_loghousesize_lm)
plot(Leverage ~ log_AvgHouseholdSize2, data = cancer_project_clear,
     xlab = "log_AvgHouseholdSize2", ylab = "Leverage")
abline(h = 4/(length(log_AvgHouseholdSize2)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_loghousesize_lm)
plot(Cooks.Dist ~ log_AvgHouseholdSize2, data = cancer_project_clear,
     xlab = "log_AvgHouseholdSize2", ylab = "Cook's distance")
abline(h = 4/(length(log_AvgHouseholdSize2)-2), col = "red")
#there are some outliers but not enough gap from the main data group
#decide not to cut the outliers

#t test for relationship
#pval = 0.2
#there is no relationship

#anova test
anova(mortal_loghousesize_lm)
#there is no relationship
```

```r
mortal_PMarried_lm <- lm(mortality~PercentMarried, data = cancer_project_clear)
summary(mortal_PMarried_lm)
plot(mortality~PercentMarried, data = cancer_project_clear)
abline(mortal_PMarried_lm)
plot(mortal_PMarried_lm)
hist(cancer_project_clear$PercentMarried)

#find outlier
Leverage <- hatvalues(mortal_PMarried_lm)
plot(Leverage ~ PercentMarried, data = cancer_project_clear,
     xlab = "PercentMarried", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$PercentMarried)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PMarried_lm)
plot(Cooks.Dist ~ PercentMarried, data = cancer_project_clear,
     xlab = "PercentMarried", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$PercentMarried)-2), col = "red")
#there are some outliers but not enough gap from the main data group
#decide not to cut the outliers

#test for relationship
#pval very low
#there might be a relationship

#anova test
#data look suit for anova
anova(mortal_PMarried_lm)
#there might be a relationship
```

```r
mortal_PctNoHS18_24_lm <- lm(mortality~PctNoHS18_24,
                             data = cancer_project_clear)
summary(mortal_PctNoHS18_24_lm)
```

```r
plot(mortality~PctNoHS18_24, data = cancer_project_clear)
abline(mortal_PctNoHS18_24_lm)
plot(mortal_PctNoHS18_24_lm)
hist(cancer_project_clear$PctNoHS18_24)

#the data look skew
#try transform the
```

```r
#try transform the data to sqrt root of PctNoHS18_24
sqrt_PctNoHS18_24 <- sqrt(cancer_project_clear$PctNoHS18_24)

mortal_sqrtPctNoHS18_24_lm <- lm(mortality~sqrt_PctNoHS18_24,
                                 data = cancer_project_clear)
summary(mortal_sqrtPctNoHS18_24_lm)
plot(mortality~sqrt_PctNoHS18_24, data = cancer_project_clear)
abline(mortal_sqrtPctNoHS18_24_lm)
plot(mortal_sqrtPctNoHS18_24_lm)
hist(sqrt_PctNoHS18_24)

#find outlier
Leverage <- hatvalues(mortal_sqrtPctNoHS18_24_lm)
plot(Leverage ~ sqrt_PctNoHS18_24, data = cancer_project_clear,
     xlab = "sqrt_PctNoHS18_24", ylab = "Leverage")
abline(h = 4/(length(sqrt_PctNoHS18_24)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_sqrtPctNoHS18_24_lm)
plot(Cooks.Dist ~ sqrt_PctNoHS18_24, data = cancer_project_clear,
     xlab = "sqrt_PctNoHS18_24", ylab = "Cook's distance")
abline(h = 4/(length(sqrt_PctNoHS18_24)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough
#decide not to cut the outliers

#test for relationship
#pval very low
#there might be a relationship

#anova test
#data look suit for anova
anova(mortal_sqrtPctNoHS18_24_lm)
#there might be a relationship
```

```r
mortal_PctHS18_24_lm <- lm(mortality~PctHS18_24, data = cancer_project_clear)
summary(mortal_PctHS18_24_lm)
plot(mortality~PctHS18_24, data = cancer_project_clear)
abline(mortal_PctHS18_24_lm)
plot(mortal_PctHS18_24_lm)
hist(cancer_project_clear$PctHS18_24)

#find outlier
Leverage <- hatvalues(mortal_PctHS18_24_lm)
plot(Leverage ~ PctHS18_24, data = cancer_project_clear,
     xlab = "PctHS18_24", ylab = "Leverage")
```

```r
abline(h = 4/(length(cancer_project_clear$PctHS18_24)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctHS18_24_lm)
plot(Cooks.Dist ~ PctHS18_24, data = cancer_project_clear,
     xlab = "PctHS18_24", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$PctHS18_24)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test for relationship
#pval very low
#there might be a relationship

#test anova
#data look suitable for anova
anova(mortal_PctHS18_24_lm)
#there might be a relationship
```

```r
mortal_PctSomeCol18_24_lm <- lm(mortality~PctSomeCol18_24,
                                data = cancer_project_clear)
summary(mortal_PctSomeCol18_24_lm)
plot(mortality~PctSomeCol18_24, data = cancer_project_clear)
abline(mortal_PctSomeCol18_24_lm)
plot(mortal_PctSomeCol18_24_lm)
hist(cancer_project_clear$PctSomeCol18_24)

#the data not have enogh to make good model
```

```r
mortal_PctBachDeg18_24_lm <- lm(mortality~PctBachDeg18_24,
                                data = cancer_project_clear)
summary(mortal_PctBachDeg18_24_lm)
plot(mortality~PctBachDeg18_24, data = cancer_project_clear)
abline(mortal_PctBachDeg18_24_lm)
plot(mortal_PctBachDeg18_24_lm)
hist(cancer_project_clear$PctBachDeg18_24)
# the data look heavy skew
#transform
```

```r
#try transform the data to 4th root of PctBachDeg18_24
forrt_PctBachDeg18_24 <- (cancer_project_clear$PctBachDeg18_24)^0.25

mortal_forrtPctBachDeg18_24_lm <- lm(mortality~forrt_PctBachDeg18_24,
                                     data = cancer_project_clear)
summary(mortal_forrtPctBachDeg18_24_lm)
plot(mortality~forrt_PctBachDeg18_24, data = cancer_project_clear)
abline(mortal_forrtPctBachDeg18_24_lm)
plot(mortal_forrtPctBachDeg18_24_lm)
hist(forrt_PctBachDeg18_24)

#find outlier
Leverage <- hatvalues(mortal_forrtPctBachDeg18_24_lm)
plot(Leverage ~ forrt_PctBachDeg18_24, data = cancer_project_clear,
```

```
     xlab = "forrt_PctBachDeg18_24", ylab = "Leverage")
abline(h = 4/(length(forrt_PctBachDeg18_24)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_forrtPctBachDeg18_24_lm)
plot(Cooks.Dist ~ forrt_PctBachDeg18_24, data = cancer_project_clear,
     xlab = "forrt_PctBachDeg18_24", ylab = "Cook's distance")
abline(h = 4/(length(forrt_PctBachDeg18_24)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test for relationship
#pval very low
#there might be a relationship

#test anova
#data look suitable for anova
anova(mortal_forrtPctBachDeg18_24_lm)
#there might be a relationship
```

```
mortal_PctHS25_Over_lm <- lm(mortality~PctHS25_Over,
                             data = cancer_project_clear)
summary(mortal_PctHS25_Over_lm)
plot(mortality~PctHS25_Over, data = cancer_project_clear)
abline(mortal_PctHS25_Over_lm)
plot(mortal_PctHS25_Over_lm)
hist(cancer_project_clear$PctHS25_Over)

#find outlier
Leverage <- hatvalues(mortal_PctHS25_Over_lm)
plot(Leverage ~ PctHS25_Over, data = cancer_project_clear,
     xlab = "PctHS25_Over", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$PctHS25_Over)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctHS25_Over_lm)
plot(Cooks.Dist ~ PctHS25_Over, data = cancer_project_clear,
     xlab = "PctHS25_Over", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$PctHS25_Over)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test for relationship
#pval is low
# their might be relationship

#anova test
#the data look suit for anova
anova(mortal_PctHS25_Over_lm)
#their might be relationship
```

```
mortal_PctBachDeg25_Over_lm <- lm(mortality~PctBachDeg25_Over,
                                  data = cancer_project_clear)
```

```r
summary(mortal_PctBachDeg25_Over_lm)
plot(mortality~PctBachDeg25_Over, data = cancer_project_clear)
abline(mortal_PctBachDeg25_Over_lm)
plot(mortal_PctBachDeg25_Over_lm)
hist(cancer_project_clear$PctBachDeg25_Over)

#data look skew
#do transform
```

```r
#try transform the data to log of PctBachDeg25_Over
log_PctBachDeg25_Over <- log(cancer_project_clear$PctBachDeg25_Over)

mortal_logPctBachDeg25_Over_lm <- lm(mortality~log_PctBachDeg25_Over,
                                     data = cancer_project_clear)
summary(mortal_logPctBachDeg25_Over_lm)
plot(mortality~log_PctBachDeg25_Over, data = cancer_project_clear)
abline(mortal_logPctBachDeg25_Over_lm)
plot(mortal_logPctBachDeg25_Over_lm)
hist(log_PctBachDeg25_Over)

#find outlier
Leverage <- hatvalues(mortal_logPctBachDeg25_Over_lm)
plot(Leverage ~ log_PctBachDeg25_Over, data = cancer_project_clear,
     xlab = "log_PctBachDeg25_Over", ylab = "Leverage")
abline(h = 4/(length(log_PctBachDeg25_Over)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_logPctBachDeg25_Over_lm)
plot(Cooks.Dist ~ log_PctBachDeg25_Over, data = cancer_project_clear,
     xlab = "log_PctBachDeg25_Over", ylab = "Cook's distance")
abline(h = 4/(length(log_PctBachDeg25_Over)-2), col = "red")
#there are some outliers and some gap from the main data
#group, but looking in the data it look valid enough.
#decide not to cut the outliers

#test for relationship
#pval is low
# their might be relationship

#anova test
#the data look suit for anova
anova(mortal_logPctBachDeg25_Over_lm)
#their might be relationship
```

```r
#cleaning data
mortality_PctEmployed16_Over <- cancer_project_clear$mortality[!is.na(
  cancer_project_clear$PctEmployed16_Over)]
PctEmployed16_Over_clean <- cancer_project_clear$PctEmployed16_Over[!is.na(
  cancer_project_clear$PctEmployed16_Over)]

mortal_PctEmployed16_Over_lm <- lm(mortality_PctEmployed16_Over~
                                     PctEmployed16_Over_clean)
summary(mortal_PctEmployed16_Over_lm)
plot(mortality_PctEmployed16_Over~PctEmployed16_Over_clean)
```

```r
abline(mortal_PctEmployed16_Over_lm)
plot(mortal_PctEmployed16_Over_lm)
hist(PctEmployed16_Over_clean)

#find outlier
Leverage <- hatvalues(mortal_PctEmployed16_Over_lm)
plot(Leverage ~ PctEmployed16_Over_clean,
     xlab = "PctEmployed16_Over", ylab = "Leverage")
abline(h = 4/(length(PctEmployed16_Over_clean)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctEmployed16_Over_lm)
plot(Cooks.Dist ~ PctEmployed16_Over_clean,
     xlab = "PctEmployed16_Over", ylab = "Cook's distance")
abline(h = 4/(length(PctEmployed16_Over_clean)-2), col = "red")
#there are some outliers and some gap from the main data
#group, but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test anova
#data look suitable for anova
anova(mortal_PctEmployed16_Over_lm)
#their might be relationship
```

```r
mortal_PctUnemployed16_Over_lm <- lm(mortality~PctUnemployed16_Over,
                                     data = cancer_project_clear)
summary(mortal_PctUnemployed16_Over_lm)
plot(mortality~PctUnemployed16_Over, data = cancer_project_clear)
abline(mortal_PctUnemployed16_Over_lm)
plot(mortal_PctUnemployed16_Over_lm)
hist(cancer_project_clear$PctUnemployed16_Over)

#data could be transfrom
```

```r
#try transform the data to sqrt of PctUnemployed16_Over
sqrt_PctUnemployed16_Over <- sqrt(cancer_project_clear$PctUnemployed16_Over)

mortal_sqrtPctUnemployed16_Over_lm <- lm(mortality~sqrt_PctUnemployed16_Over,
                                         data = cancer_project_clear)
summary(mortal_sqrtPctUnemployed16_Over_lm)
plot(mortality~sqrt_PctUnemployed16_Over, data = cancer_project_clear)
abline(mortal_sqrtPctUnemployed16_Over_lm)
plot(mortal_sqrtPctUnemployed16_Over_lm)
hist(sqrt_PctUnemployed16_Over)

#find outlier
Leverage <- hatvalues(mortal_sqrtPctUnemployed16_Over_lm)
plot(Leverage ~ sqrt_PctUnemployed16_Over,
     xlab = "sqrt_PctUnemployed16_Over", ylab = "Leverage")
abline(h = 4/(length(sqrt_PctUnemployed16_Over)-2), col = "red")
```

```r
Cooks.Dist <- cooks.distance(mortal_sqrtPctUnemployed16_Over_lm)
plot(Cooks.Dist ~ sqrt_PctUnemployed16_Over,
     xlab = "sqrt_PctUnemployed16_Over", ylab = "Cook's distance")
abline(h = 4/(length(sqrt_PctUnemployed16_Over)-2), col = "red")
#there are some outliers and huge gap from the main data
#group,but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test anova
#data look suitable for anova
anova(mortal_sqrtPctUnemployed16_Over_lm)
#their might be relationship
```

```r
mortal_PctPrivateCoverage_lm <- lm(mortality~PctPrivateCoverage,
                                   data = cancer_project_clear)
summary(mortal_PctPrivateCoverage_lm)
plot(mortality~PctPrivateCoverage, data = cancer_project_clear)
abline(mortal_PctPrivateCoverage_lm)
plot(mortal_PctPrivateCoverage_lm)
hist(cancer_project_clear$PctPrivateCoverage)
#there is higher variance at the high value of PctPrivateCoverage
#might due to lack of data point

#find outlier
Leverage <- hatvalues(mortal_PctPrivateCoverage_lm)
plot(Leverage ~ PctPrivateCoverage, data = cancer_project_clear,
     xlab = "PctPrivateCoverage", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$PctPrivateCoverage)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctPrivateCoverage_lm)
plot(Cooks.Dist ~ PctPrivateCoverage, data = cancer_project_clear,
     xlab = "PctPrivateCoverage", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$PctPrivateCoverage)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test for anova
#data look suitable for anova test
anova(mortal_PctPrivateCoverage_lm)
#their might be relation ship
```

```r
#cleaning data
mortality_PctPrivateCoverageAlone <- cancer_project_clear$mortality[
  !is.na(cancer_project_clear$PctPrivateCoverageAlone)]
```

```r
PctPrivateCoverageAlone_clean <- cancer_project_clear$PctPrivateCoverageAlone[
  !is.na(cancer_project_clear$PctPrivateCoverageAlone)]

mortal_PctPrivateCoverageAlone_lm <- lm(mortality_PctPrivateCoverageAlone
                                        ~PctPrivateCoverageAlone_clean)
summary(mortal_PctPrivateCoverageAlone_lm)
plot(mortality_PctPrivateCoverageAlone~PctPrivateCoverageAlone_clean)
abline(mortal_PctPrivateCoverageAlone_lm)
plot(mortal_PctPrivateCoverageAlone_lm)
#there is higher variance at the high value of PctPrivateCoverageAlone
#might due to lack of data point

#find outlier
Leverage <- hatvalues(mortal_PctPrivateCoverageAlone_lm)
plot(Leverage ~ PctPrivateCoverageAlone_clean ,
     data = cancer_project_clear,
     xlab = "PctPrivateCoverageAlone", ylab = "Leverage")
abline(h = 4/(length(PctPrivateCoverageAlone_clean)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctPrivateCoverageAlone_lm)
plot(Cooks.Dist ~ PctPrivateCoverageAlone_clean,
     data = cancer_project_clear,
     xlab = "PctPrivateCoverageAlone", ylab = "Cook's distance")
abline(h = 4/(length(PctPrivateCoverageAlone_clean)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test for anova
hist(cancer_project_clear$PctPrivateCoverageAlone)
#data look suitable for anova
anova(mortal_PctPrivateCoverageAlone_lm)
#their might be relationship
```

```r
#cleaning data
mortality_PctEmpPrivCoverage <- cancer_project_clear$mortality[
  !is.na(cancer_project_clear$PctEmpPrivCoverage)]
PctEmpPrivCoverage_clean <- cancer_project_clear$PctEmpPrivCoverage[
  !is.na(cancer_project_clear$PctEmpPrivCoverage)]

mortal_PctEmpPrivCoverage_lm <- lm(mortality_PctEmpPrivCoverage
                                   ~PctEmpPrivCoverage_clean)
summary(mortal_PctEmpPrivCoverage_lm)
plot(mortality_PctEmpPrivCoverage~PctEmpPrivCoverage_clean)
abline(mortal_PctEmpPrivCoverage_lm)
plot(mortal_PctEmpPrivCoverage_lm)
hist(PctEmpPrivCoverage_clean)
#there is higher variance at the high value of PctEmpPrivCoverage
#might due to lack of data point
```

```r
#find outlier
Leverage <- hatvalues(mortal_PctEmpPrivCoverage_lm)
plot(Leverage ~ PctEmpPrivCoverage_clean ,data = cancer_project_clear,
     xlab = "PctEmpPrivCoverage", ylab = "Leverage")
abline(h = 4/(length(PctEmpPrivCoverage_clean)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctEmpPrivCoverage_lm)
plot(Cooks.Dist ~ PctEmpPrivCoverage_clean, data = cancer_project_clear,
     xlab = "PctEmpPrivCoverage", ylab = "Cook's distance")
abline(h = 4/(length(PctEmpPrivCoverage_clean)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#anova test
#the data look suit fot anvoa test
anova(mortal_PctEmpPrivCoverage_lm)
#their might be relationship
```

```r
mortal_PctPublicCoverage_lm <- lm(mortality~PctPublicCoverage,
                                  data = cancer_project_clear)
summary(mortal_PctPublicCoverage_lm)
plot(mortality~PctPublicCoverage, data = cancer_project_clear)
abline(mortal_PctPublicCoverage_lm)
plot(mortal_PctPublicCoverage_lm)

#find outlier
Leverage <- hatvalues(mortal_PctPublicCoverage_lm)
plot(Leverage ~ PctPublicCoverage , data = cancer_project_clear,
     xlab = "PctPublicCoverage", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$PctPublicCoverage)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctPublicCoverage_lm)
plot(Cooks.Dist ~ PctPublicCoverage, data = cancer_project_clear,
     xlab = "PctPublicCoverage", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$PctPublicCoverage)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#anova test
#the data look suit fot anvoa test
anova(mortal_PctPublicCoverage_lm)
#their might be relationship
```

```r
mortal_PctPublicCoverageAlone_lm <- lm(mortality~PctPublicCoverageAlone,
                                       data = cancer_project_clear)
summary(mortal_PctPublicCoverageAlone_lm)
plot(mortality~PctPublicCoverageAlone, data = cancer_project_clear)
abline(mortal_PctPublicCoverageAlone_lm)
plot(mortal_PctPublicCoverageAlone_lm)
hist(cancer_project_clear$PctPublicCoverageAlone)

#find outlier
Leverage <- hatvalues(mortal_PctPublicCoverageAlone_lm)
plot(Leverage ~ PctPublicCoverageAlone , data = cancer_project_clear,
     xlab = "PctPublicCoverageAlone", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$PctPublicCoverageAlone)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_PctPublicCoverageAlone_lm)
plot(Cooks.Dist ~ PctPublicCoverageAlone, data = cancer_project_clear,
     xlab = "PctPublicCoverageAlone", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$PctPublicCoverageAlone)-2), col = "red")
#there are some outliers and some gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#anova test
#the data look suit fot anvoa test
anova(mortal_PctPublicCoverageAlone_lm)
#their might be relationship
```

```r
mortal_PctWhite_lm <- lm(mortality~PctWhite,
                                       data = cancer_project_clear)
summary(mortal_PctWhite_lm)
plot(mortality~PctWhite, data = cancer_project_clear)
abline(mortal_PctWhite_lm)
plot(mortal_PctWhite_lm)
hist(cancer_project_clear$PctWhite)
#needed transformation
```

```r
#try transform the data to 4th power of PctWhite
forpo_PctWhite <- (cancer_project_clear$PctWhite)^4

mortal_forpoPctWhite_lm <- lm(mortality~forpo_PctWhite,
                              data = cancer_project_clear)
summary(mortal_forpoPctWhite_lm)
plot(mortality~forpo_PctWhite, data = cancer_project_clear)
abline(mortal_forpoPctWhite_lm)
plot(mortal_forpoPctWhite_lm)
hist(forpo_PctWhite)

#find outlier
Leverage <- hatvalues(mortal_forpoPctWhite_lm)
```

```r
plot(Leverage ~ forpo_PctWhite, xlab = "forpo_PctWhite", ylab = "Leverage")
abline(h = 4/(length(forpo_PctWhite)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_forpoPctWhite_lm)
plot(Cooks.Dist ~ forpo_PctWhite,
     xlab = "forpo_PctWhite", ylab = "Cook's distance")
abline(h = 4/(length(forpo_PctWhite)-2), col = "red")
#there are some outliers and huge gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test anova
#data might not suitable for anova
anova(mortal_forpoPctWhite_lm)
#their might be relationship
```

```r
mortal_PctBlack_lm <- lm(mortality~PctBlack,
                                      data = cancer_project_clear)
summary(mortal_PctBlack_lm)
plot(mortality~PctBlack, data = cancer_project_clear)
abline(mortal_PctBlack_lm)
plot(mortal_PctBlack_lm)
hist(cancer_project_clear$PctBlack)

#needed transformation
```

```r
#try transform the data to 4th root of PctWhite
forrt_PctBlack <- (cancer_project_clear$PctBlack)^0.25

mortal_forrtPctBlack_lm <- lm(mortality~forrt_PctBlack,
                              data = cancer_project_clear)
summary(mortal_forrtPctBlack_lm)
plot(mortality~forrt_PctBlack, data = cancer_project_clear)
abline(mortal_forrtPctBlack_lm)
plot(mortal_forrtPctBlack_lm)
hist(forrt_PctBlack)

#find outlier
Leverage <- hatvalues(mortal_forrtPctBlack_lm)
plot(Leverage ~ forrt_PctBlack, xlab = "forrt_PctBlack", ylab = "Leverage")
abline(h = 4/(length(forrt_PctBlack)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_forrtPctBlack_lm)
plot(Cooks.Dist ~ forrt_PctBlack,
     xlab = "forrt_PctBlack", ylab = "Cook's distance")
abline(h = 4/(length(forrt_PctBlack)-2), col = "red")
#there are some outliers and huge gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers
```

```r
#test statistic
#pval is low
#their might be relationship

#test anova
#data might not suitable for anova
anova(mortal_forrtPctBlack_lm)
#their might be relationship
```

```r
mortal_PctAsian_lm <- lm(mortality~PctAsian,
                                        data = cancer_project_clear)
summary(mortal_PctAsian_lm)
plot(mortality~PctAsian, data = cancer_project_clear)
abline(mortal_PctAsian_lm)
plot(mortal_PctAsian_lm)

#needed transformation
```

```r
#try transform the data to 5th root of PctWhite
fifrt_PctAsian <- (cancer_project_clear$PctAsian)^0.2

mortal_fifrtPctAsian_lm <- lm(mortality~fifrt_PctAsian, data = cancer_project_clear)
summary(mortal_fifrtPctAsian_lm)
plot(mortality~fifrt_PctAsian, data = cancer_project_clear)
abline(mortal_fifrtPctAsian_lm)
plot(mortal_fifrtPctAsian_lm)
hist(fifrt_PctAsian)

#find outlier
Leverage <- hatvalues(mortal_fifrtPctAsian_lm)
plot(Leverage ~ fifrt_PctAsian, xlab = "fifrt_PctAsian", ylab = "Leverage")
abline(h = 4/(length(fifrt_PctAsian)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_fifrtPctAsian_lm)
plot(Cooks.Dist ~ fifrt_PctAsian,
     xlab = "fifrt_PctAsian", ylab = "Cook's distance")
abline(h = 4/(length(fifrt_PctAsian)-2), col = "red")
#there are some outliers and huge gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test anova
#data might not suitable for anova
anova(mortal_fifrtPctAsian_lm)
#their might be relationship
```

```r
mortal_PctOtherRace_lm <- lm(mortality~PctOtherRace,
                                        data = cancer_project_clear)
```

```r
summary(mortal_PctOtherRace_lm)
plot(mortality~PctOtherRace, data = cancer_project_clear)
abline(mortal_PctOtherRace_lm)
plot(mortal_PctOtherRace_lm)

#needed transformation
```

```r
#try transform the data to 5th root of PctWhite
fifrt_PctOtherRace <- (cancer_project_clear$PctOtherRace)^0.2

mortal_fifrtPctOtherRace_lm <- lm(mortality~fifrt_PctOtherRace, data = cancer_project_clear)
summary(mortal_fifrtPctOtherRace_lm)
plot(mortality~fifrt_PctOtherRace, data = cancer_project_clear)
abline(mortal_fifrtPctOtherRace_lm)
plot(mortal_fifrtPctOtherRace_lm)
hist(fifrt_PctOtherRace)

#find outlier
Leverage <- hatvalues(mortal_fifrtPctOtherRace_lm)
plot(Leverage ~ fifrt_PctOtherRace,
     xlab = "fifrt_PctOtherRace", ylab = "Leverage")
abline(h = 4/(length(fifrt_PctOtherRace)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_fifrtPctOtherRace_lm)
plot(Cooks.Dist ~ fifrt_PctOtherRace,
     xlab = "fifrt_PctOtherRace", ylab = "Cook's distance")
abline(h = 4/(length(fifrt_PctOtherRace)-2), col = "red")
#there are some outliers and huge gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test anova
#data might not suitable for anova
anova(mortal_fifrtPctOtherRace_lm)
#their might be relationship
```

```r
mortal_PctMarriedHouseholds_lm <- lm(mortality~PctMarriedHouseholds,
                                     data = cancer_project_clear)
summary(mortal_PctMarriedHouseholds_lm)
plot(mortality~PctMarriedHouseholds, data = cancer_project_clear)
abline(mortal_PctMarriedHouseholds_lm)
plot(mortal_PctMarriedHouseholds_lm)
hist(cancer_project_clear$PctMarriedHouseholds)

#find outlier
Leverage <- hatvalues(mortal_PctMarriedHouseholds_lm)
plot(Leverage ~ PctMarriedHouseholds, data = cancer_project_clear,
     xlab = "PctMarriedHouseholds", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$PctMarriedHouseholds)-2), col = "red")
```

```r
Cooks.Dist <- cooks.distance(mortal_PctMarriedHouseholds_lm)
plot(Cooks.Dist ~ fifrt_PctOtherRace, data = cancer_project_clear,
     xlab = "PctMarriedHouseholds", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$PctMarriedHouseholds)-2), col = "red")
#there are some outliers and huge gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
#their might be relationship

#test anova
#data might not suitable for anova
anova(mortal_PctMarriedHouseholds_lm)
#their might be relationship
```

```r
mortal_BirthRate_lm <- lm(mortality~BirthRate,
                                    data = cancer_project_clear)
summary(mortal_BirthRate_lm)
plot(mortality~BirthRate, data = cancer_project_clear)
abline(mortal_BirthRate_lm)
plot(mortal_BirthRate_lm)
hist(cancer_project_clear$BirthRate)

#need transform
```

```r
#try transform the data to sqrt of BirthRate
sqrt_BirthRate <- (cancer_project_clear$BirthRate)^0.5

mortal_sqrtBirthRate_lm <- lm(mortality~sqrt_BirthRate,
                              data = cancer_project_clear)
summary(mortal_sqrtBirthRate_lm)
plot(mortality~sqrt_BirthRate, data = cancer_project_clear)
abline(mortal_sqrtBirthRate_lm)
plot(mortal_sqrtBirthRate_lm)
hist(sqrt_BirthRate)

#find outlier
Leverage <- hatvalues(mortal_sqrtBirthRate_lm)
plot(Leverage ~ sqrt_BirthRate, xlab = "sqrt_BirthRate", ylab = "Leverage")
abline(h = 4/(length(sqrt_BirthRate)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_sqrtBirthRate_lm)
plot(Cooks.Dist ~ sqrt_BirthRate,
     xlab = "sqrt_BirthRate", ylab = "Cook's distance")
abline(h = 4/(length(sqrt_BirthRate)-2), col = "red")
#there are some outliers and huge gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test statistic
#pval is low
```

```r
#their might be relationship

#test anova
#data might not suitable for anova
anova(mortal_sqrtBirthRate_lm)
#their might be relationship
```

```r
mortal_BinnedInc_lm <- lm(mortality~BinnedInc,
                                        data = cancer_project_clear)
summary(mortal_BinnedInc_lm)
plot(mortality~BinnedInc, data = cancer_project_clear)
abline(mortal_BinnedInc_lm)
plot(mortal_BinnedInc_lm)
hist(cancer_project_clear$BinnedInc)

#find outlier
Leverage <- hatvalues(mortal_BinnedInc_lm)
plot(Leverage ~ BinnedInc, data = cancer_project_clear,
     xlab = "BinnedInc", ylab = "Leverage")
abline(h = 4/(length(cancer_project_clear$BinnedInc)-2), col = "red")

Cooks.Dist <- cooks.distance(mortal_BinnedInc_lm)
plot(Cooks.Dist ~ BinnedInc, data = cancer_project_clear,
     xlab = "BinnedInc", ylab = "Cook's distance")
abline(h = 4/(length(cancer_project_clear$BinnedInc)-2), col = "red")
#there are some outliers and huge gap from the main data group,
#but looking in the data it look valid enough.
#decide not to cut the outliers

#test for relationship
#pval is low
#there might be a relationship

#anova test
#the data not suit for anova
anova(mortal_BinnedInc_lm)
#the model might not be a good model in the first place
```

```r
#list of all final model at have relationship
SLR <- data.frame(model = c(
                           "mortal_logmedIncome_lm",
                           "mortal_logpopEst2015_lm",
                           "mortal_sqrtpovertyPercent_lm",
                           "mortal_forrtstudyPerCap_lm",
                           "mortal_PMarried_lm",
                           "mortal_sqrtPctNoHS18_24_lm",
                           "mortal_PctHS18_24_lm",
                           "mortal_forrtPctBachDeg18_24_lm",
                           "mortal_PctHS25_Over_lm",
                           "mortal_logPctBachDeg25_Over_lm",
                           "mortal_PctEmployed16_Over_lm",
                           "mortal_sqrtPctUnemployed16_Over_lm",
                           "mortal_PctPrivateCoverage_lm",
```

```
                "mortal_PctPrivateCoverageAlone_lm",
                "mortal_PctEmpPrivCoverage_lm",
                "mortal_PctPublicCoverage_lm",
                "mortal_PctPublicCoverageAlone_lm",
                "mortal_forpoPctWhite_lm",
                "mortal_forrtPctBlack_lm",
                "mortal_fifrtPctAsian_lm",
                "mortal_fifrtPctOtherRace_lm",
                "mortal_PctMarriedHouseholds_lm",
                "mortal_sqrtBirthRate_lm",
                "mortal_BinnedInc_lm"),
    R_sq = c(
                summary(mortal_logmedIncome_lm)$r.squared,
                summary(mortal_logpopEst2015_lm)$r.squared,
                summary(mortal_sqrtpovertyPercent_lm)$r.squared,
                summary(mortal_forrtstudyPerCap_lm)$r.squared,
                summary(mortal_PMarried_lm)$r.squared,
                summary(mortal_sqrtPctNoHS18_24_lm)$r.squared,
                summary(mortal_PctHS18_24_lm)$r.squared,
                summary(mortal_forrtPctBachDeg18_24_lm)$r.squared,
                summary(mortal_PctHS25_Over_lm)$r.squared,
                summary(mortal_logPctBachDeg25_Over_lm)$r.squared,
                summary(mortal_PctEmployed16_Over_lm)$r.squared,
                summary(mortal_sqrtPctUnemployed16_Over_lm)$r.squared,
                summary(mortal_PctPrivateCoverage_lm)$r.squared,
                summary(mortal_PctPrivateCoverageAlone_lm)$r.squared,
                summary(mortal_PctEmpPrivCoverage_lm)$r.squared,
                summary(mortal_PctPublicCoverage_lm)$r.squared,
                summary(mortal_PctPublicCoverageAlone_lm)$r.squared,
                summary(mortal_forpoPctWhite_lm)$r.squared,
                summary(mortal_forrtPctBlack_lm)$r.squared,
                summary(mortal_fifrtPctAsian_lm)$r.squared,
                summary(mortal_fifrtPctOtherRace_lm)$r.squared,
                summary(mortal_PctMarriedHouseholds_lm)$r.squared,
                summary(mortal_sqrtBirthRate_lm)$r.squared,
                summary(mortal_BinnedInc_lm)$r.squared),
    MSE = c(
               mean(summary(
                 mortal_logmedIncome_lm)$residual^2),
               mean(summary(
                 mortal_logpopEst2015_lm)$residual^2),
               mean(summary(
                 mortal_sqrtpovertyPercent_lm)$residual^2),
               mean(summary(
                 mortal_forrtstudyPerCap_lm)$residual^2),
               mean(summary(
                 mortal_PMarried_lm)$residual^2),
               mean(summary(
                 mortal_sqrtPctNoHS18_24_lm)$residual^2),
               mean(summary(
                 mortal_PctHS18_24_lm)$residual^2),
               mean(summary(
                 mortal_forrtPctBachDeg18_24_lm)$residual^2),
```

```
                          mean(summary(
                            mortal_PctHS25_Over_lm)$residual^2),
                          mean(summary(
                            mortal_logPctBachDeg25_Over_lm)$residual^2),
                          mean(summary(
                            mortal_PctEmployed16_Over_lm)$residual^2),
                          mean(summary(
                            mortal_sqrtPctUnemployed16_Over_lm)$residual^2),
                          mean(summary(
                            mortal_PctPrivateCoverage_lm)$residual^2),
                          mean(summary(
                            mortal_PctPrivateCoverageAlone_lm)$residual^2),
                          mean(summary(
                            mortal_PctEmpPrivCoverage_lm)$residual^2),
                          mean(summary(
                            mortal_PctPublicCoverage_lm)$residual^2),
                          mean(summary(
                            mortal_PctPublicCoverageAlone_lm)$residual^2),
                          mean(summary(
                            mortal_forpoPctWhite_lm)$residual^2),
                          mean(summary(
                            mortal_forrtPctBlack_lm)$residual^2),
                          mean(summary(
                            mortal_fifrtPctAsian_lm)$residual^2),
                          mean(summary(
                            mortal_fifrtPctOtherRace_lm)$residual^2),
                          mean(summary(
                            mortal_PctMarriedHouseholds_lm)$residual^2),
                          mean(summary(
                            mortal_sqrtBirthRate_lm)$residual^2),
                          mean(summary(
                            mortal_BinnedInc_lm)$residual^2)
                        )
                    )
View(SLR)
#the best model
SLR$model[SLR$R_sq == max(SLR$R_sq)]
SLR$model[SLR$MSE == min(SLR$MSE)]
```

```
#graph for report
par(mfrow = c(2, 2))
plot(mortal_logPctBachDeg25_Over_lm)
```

```
#MLR
#remove geo
#remove PctSomeCol18_24 due to low data point
cancer_project_remove <- cancer_project_clear[-c(12,17)]
#remove NA datapoint from PctEmployed16_Over,PctPrivateCoverageAlone,PctEmpPrivCoverage
cancer_project_na1 <- cancer_project_remove[
  !is.na(cancer_project_remove$PctEmployed16_Over),]
cancer_project_na2 <- cancer_project_na1[
  !is.na(cancer_project_na1$PctPrivateCoverageAlone),]
cancer_project_na3 <- cancer_project_na2[
```

```
  !is.na(cancer_project_na2$PctEmpPrivCoverage),]

#remove outlier
#remove < 1 data point
#AvgHouseholdSize
cancer_project_out1 <- cancer_project_na3[
  cancer_project_na3$AvgHouseholdSize >= 1,]

#remove >200 medage
cancer_project_out2 <- cancer_project_out1[cancer_project_out1$MedianAge <200,]

cancer_project_remove <- cancer_project_out2
```

```
#transfrom data
#log_incidenceRate


#log_medIncome
cancer_project_remove_medIncome <- cancer_project_remove[-c(5)]
cancer_project_transform_medIncome <- cbind(cancer_project_remove_medIncome,
                                            log_medIncome = log(
                                              cancer_project_remove$medIncome))

#log_popEst2015
cancer_project_remove_popEst2015 <- cancer_project_transform_medIncome[-c(5)]
cancer_project_transform_popEst2015 <-
  cbind(cancer_project_remove_popEst2015,
        log_popEst2015 = log(cancer_project_remove$popEst2015))

#sqrt_povertyPercent
cancer_project_remove_povertyPercent <-
  cancer_project_transform_popEst2015[-c(5)]
cancer_project_transform_povertyPercent <-
  cbind(cancer_project_remove_povertyPercent,
        sqrt_povertyPercent = sqrt(cancer_project_remove$povertyPercent))

#forrt_studyPerCap
cancer_project_remove_studyPerCap <-
  cancer_project_transform_povertyPercent[-c(5)]
cancer_project_transform_studyPerCap <-
  cbind(cancer_project_remove_studyPerCap,
        forrt_studyPerCap = (cancer_project_remove$studyPerCap)^0.25)

#log_AvgHouseholdSize
cancer_project_remove_AvgHouseholdSize <-
  cancer_project_transform_studyPerCap[-c(8)]
cancer_project_transform_AvgHouseholdSize <-
  cbind(cancer_project_remove_AvgHouseholdSize,
        log_AvgHouseholdSize = log(cancer_project_remove$AvgHouseholdSize))

#sqrt_PctNoHS18_24
cancer_project_remove_PctNoHS18_24 <-
  cancer_project_transform_AvgHouseholdSize[-c(9)]
```

```r
cancer_project_transform_PctNoHS18_24 <-
  cbind(cancer_project_remove_PctNoHS18_24,
        sqrt_PctNoHS18_24 = sqrt(cancer_project_remove$PctNoHS18_24))

#forrt_PctBachDeg18_24
cancer_project_remove_PctBachDeg18_24 <-
  cancer_project_transform_PctNoHS18_24[-c(10)]
cancer_project_transform_PctBachDeg18_24 <-
  cbind(cancer_project_remove_PctBachDeg18_24 ,
        forrt_PctBachDeg18_24 = (cancer_project_remove$PctBachDeg18_24)^0.25)

#log_PctBachDeg25_Over
cancer_project_remove_PctBachDeg25_Over <-
  cancer_project_transform_PctBachDeg18_24[-c(11)]
cancer_project_transform_PctBachDeg25_Over <-
  cbind(cancer_project_remove_PctBachDeg25_Over,
        log_PctBachDeg25_Over = log(cancer_project_remove$PctBachDeg25_Over))

#sqrt_PctUnemployed16_Over
cancer_project_remove_PctUnemployed16_Over <-
  cancer_project_transform_PctBachDeg25_Over[-c(12)]
cancer_project_transform_PctUnemployed16_Over <-
  cbind(cancer_project_remove_PctUnemployed16_Over,
        sqrt_PctUnemployed16_Over = sqrt(
          cancer_project_remove$PctUnemployed16_Over))

#forpo_PctWhite
cancer_project_remove_PctWhite <-
  cancer_project_transform_PctUnemployed16_Over[-c(17)]
cancer_project_transform_PctWhite <-
  cbind(cancer_project_remove_PctWhite,
        forpo_PctWhite = (cancer_project_remove$PctWhite)^4)

#forrt_PctBlack
cancer_project_remove_PctBlack <- cancer_project_transform_PctWhite[-c(17)]
cancer_project_transform_PctBlack <-
  cbind(cancer_project_remove_PctBlack,
        forrt_PctBlack = (cancer_project_remove$PctBlack)^0.25)

#fifrt_PctAsian
cancer_project_remove_PctAsian <- cancer_project_transform_PctBlack[-c(17)]
cancer_project_transform_PctAsian <-
  cbind(cancer_project_remove_PctAsian,
        fifrt_PctAsian = (cancer_project_remove$PctAsian)^0.2)

#fifrt_PctOtherRace
cancer_project_remove_PctOtherRace <- cancer_project_transform_PctAsian[-c(17)]
cancer_project_transform_PctOtherRace <-
  cbind(cancer_project_remove_PctOtherRace,
        fifrt_PctOtherRace = (cancer_project_remove$PctOtherRace)^0.2)

#sqrt_BirthRate
cancer_project_remove_BirthRate <- cancer_project_transform_PctOtherRace[-c(18)]
```

```r
cancer_project_transform_BirthRate <-
  cbind(cancer_project_remove_BirthRate,
        sqrt_BirthRate = sqrt(cancer_project_remove$BirthRate))

cancer_project_MLR <- cancer_project_transform_BirthRate
View(cancer_project_MLR)
```

```r
#since there collation betaween anncount and death per year remove them
cancer_project_MLR2 <- cancer_project_MLR[-c(1,2,4)]
View(cancer_project_MLR2)
library(PerformanceAnalytics)
chart.Correlation(cancer_project_MLR[, -28])
```

```r
#using forward sequential
lm.0 <- lm(mortality ~ 1, data = cancer_project_MLR2)
lm.forward <- step(lm.0, scope = ~ MedianAge +
                             MedianAgeMale +
                             MedianAgeFemale +
                             PercentMarried +
                             PctHS18_24 +
                             PctHS25_Over +
                             PctEmployed16_Over +
                             PctPrivateCoverage +
                             PctPrivateCoverageAlone +
                             PctEmpPrivCoverage +
                             PctPublicCoverage +
                             PctPublicCoverageAlone +
                             PctMarriedHouseholds +
                             BinnedInc +

                             log_medIncome +
                             log_popEst2015 +
                             sqrt_povertyPercent +
                             forrt_studyPerCap +
                             log_AvgHouseholdSize +
                             sqrt_PctNoHS18_24 +
                             forrt_PctBachDeg18_24 +
                             log_PctBachDeg25_Over +
                             sqrt_PctUnemployed16_Over +
                             forpo_PctWhite +
                             forrt_PctBlack +
                             fifrt_PctAsian +
                             fifrt_PctOtherRace +
                             sqrt_BirthRate,
                   direction = "forward")
```

```r
summary(lm(mortality ~ log_PctBachDeg25_Over + PctMarriedHouseholds +
             PctHS25_Over + forrt_PctBlack + fifrt_PctOtherRace +
             log_popEst2015 + sqrt_BirthRate +
    PctHS18_24 + MedianAgeFemale + PercentMarried + sqrt_PctUnemployed16_Over +
    forrt_PctBachDeg18_24 + sqrt_PctNoHS18_24 + PctPublicCoverageAlone +
    PctEmpPrivCoverage + PctEmployed16_Over + PctPrivateCoverageAlone ,
    data = cancer_project_MLR2))
```

```
#comapairing final model with model before adding log_medIncome
lm.1 <- lm(mortality ~ log_PctBachDeg25_Over + PctMarriedHouseholds +
             PctHS25_Over + forrt_PctBlack + fifrt_PctOtherRace +
             log_popEst2015 + sqrt_BirthRate + PctHS18_24 + MedianAgeFemale
           + PercentMarried + sqrt_PctUnemployed16_Over + forrt_PctBachDeg18_24
           + sqrt_PctNoHS18_24 + PctPublicCoverageAlone + PctEmpPrivCoverage +
             PctEmployed16_Over + PctPrivateCoverageAlone ,
           data = cancer_project_MLR2)

lm.test <- lm(mortality ~ log_PctBachDeg25_Over + PctMarriedHouseholds +
                PctHS25_Over + forrt_PctBlack + fifrt_PctOtherRace +
                log_popEst2015 + sqrt_BirthRate + PctHS18_24 + MedianAgeFemale +
                PercentMarried + sqrt_PctUnemployed16_Over +
                forrt_PctBachDeg18_24 + sqrt_PctNoHS18_24 +
                PctPublicCoverageAlone + PctEmpPrivCoverage + PctEmployed16_Over
              , data = cancer_project_MLR2)

anova(lm.test,lm.1)
#their is weak evidence to keeping the PctPublicCoverageAlone
# decide to keep them
```

```
#remove sqrt_PctUnemployed16_Over
lm.test <- lm(mortality ~ log_PctBachDeg25_Over + PctMarriedHouseholds +
                PctHS25_Over + forrt_PctBlack + fifrt_PctOtherRace +
                log_popEst2015 + sqrt_BirthRate + PctHS18_24 + MedianAgeFemale +
                PercentMarried +  forrt_PctBachDeg18_24 + sqrt_PctNoHS18_24 +
                PctPublicCoverageAlone + PctEmpPrivCoverage + PctEmployed16_Over
              + PctPrivateCoverageAlone ,data = cancer_project_MLR2)
anova(lm.test,lm.1)
#their high chance that model 1 and test are the same
#sqrt_PctUnemployed16_Over the value shoudn't been add to the model
lm.1 <- lm.test
```

```
lm.1 <- lm.test
summary(lm.1)
```

```
# try remove PctPublicCoverageAlone
lm.test <- lm(mortality ~ log_PctBachDeg25_Over + PctMarriedHouseholds +
                PctHS25_Over +
                forrt_PctBlack + fifrt_PctOtherRace + log_popEst2015
              + sqrt_BirthRate +
                PctHS18_24 + MedianAgeFemale + PercentMarried +
                forrt_PctBachDeg18_24 + sqrt_PctNoHS18_24 +
                PctEmpPrivCoverage + PctEmployed16_Over +
                PctPrivateCoverageAlone ,data = cancer_project_MLR2)
anova(lm.test,lm.1)
#their high chance that model 1 and test are the same
#PctPublicCoverageAlone the value shoudn't been add to the model
```

```
lm.1 <- lm.test
summary(lm.1)
```

```r
# try remove PctHS18_24
lm.test <- lm(mortality ~ log_PctBachDeg25_Over + PctMarriedHouseholds +
                PctHS25_Over +
                forrt_PctBlack + fifrt_PctOtherRace + log_popEst2015 +
                sqrt_BirthRate +
                 MedianAgeFemale + PercentMarried +
                forrt_PctBachDeg18_24 + sqrt_PctNoHS18_24 +
                PctEmpPrivCoverage + PctEmployed16_Over + PctPrivateCoverageAlone,
                data = cancer_project_MLR2)
anova(lm.test,lm.1)
#their a chance that model 1 and test are not the same with 95 percent confidence interval
#PctHS18_24  should be in the model
```

```r
#Perform diagnostics checking of the model
par(mfrow = c(1, 2))
qqnorm(rstandard(lm.1))
qqline(rstandard(lm.1))
plot(lm.1$fitted.values, rstandard(lm.1),
     xlab = "Fitted Values", ylab = "Standardised Residuals")
```

```r
par(mfrow = c(3, 5))
par(mar = c(5, 4, 1, 1) + 0.1)
plot(cancer_project_MLR2$log_PctBachDeg25_Over, rstandard(lm.1),
     xlab = "log_PctBachDeg25_Over", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$PctMarriedHouseholds, rstandard(lm.1),
     xlab = "PctMarriedHouseholds", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$PctHS25_Over, rstandard(lm.1),
     xlab = "PctHS25_Over", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$forrt_PctBlack, rstandard(lm.1),
     xlab = "forrt_PctBlack", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$fifrt_PctOtherRace, rstandard(lm.1),
     xlab = "fifrt_PctOtherRace", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$log_popEst2015, rstandard(lm.1),
     xlab = "log_popEst2015", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$sqrt_BirthRate, rstandard(lm.1),
     xlab = "sqrt_BirthRate", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$PctHS18_24, rstandard(lm.1),
     xlab = "PctHS18_24", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$MedianAgeFemale, rstandard(lm.1),
     xlab = "MedianAgeFemale", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$PercentMarried, rstandard(lm.1),
     xlab = "PercentMarried", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$forrt_PctBachDeg18_24, rstandard(lm.1),
     xlab = "forrt_PctBachDeg18_24", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$sqrt_PctNoHS18_24, rstandard(lm.1),
     xlab = "sqrt_PctNoHS18_24", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$PctEmpPrivCoverage, rstandard(lm.1),
     xlab = "PctEmpPrivCoverage", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$PctEmployed16_Over, rstandard(lm.1),
     xlab = "PctEmployed16_Over", ylab = "Standardised Residuals")
plot(cancer_project_MLR2$PctPrivateCoverageAlone, rstandard(lm.1),
     xlab = "PctPrivateCoverageAlone", ylab = "Standardised Residuals")
```

```r
par(mfrow = c(2, 2))
plot(lm.1)
```

```r
Leverage <- hatvalues(lm.1)
plot(rstandard(lm.1) ~ Leverage)
abline(v = 2*(29+1)/1482)

#there are some outliers

Cooks.Dist <- cooks.distance(lm.1)
plot(rstandard(lm.1) ~ Cooks.Dist)
abline(v= 2*(29+1)/(1482-29+1))
#there are one influential point
```

```r
plot(cooks.distance(lm.1), xlab = "motality", ylab = "Cook's distance")
abline(h = 2*(28+1)/(1482-28+1))
with(cancer_project_MLR2, text(cooks.distance(lm.1),
                               labels = row.names(cancer_project_MLR2), pos = 4))
# still I want the model to consider this point of data as well
```