ST404 Small Report

Alex

03/02/2022

ST404 Project

First we load the dataset and required libraries

```
load("cancer.rdata")
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(usmap)
library(car)
```

```
## Loading required package: carData
```

```
library(stringr)
```

Summary

First lets have a look at a summary of all the data:

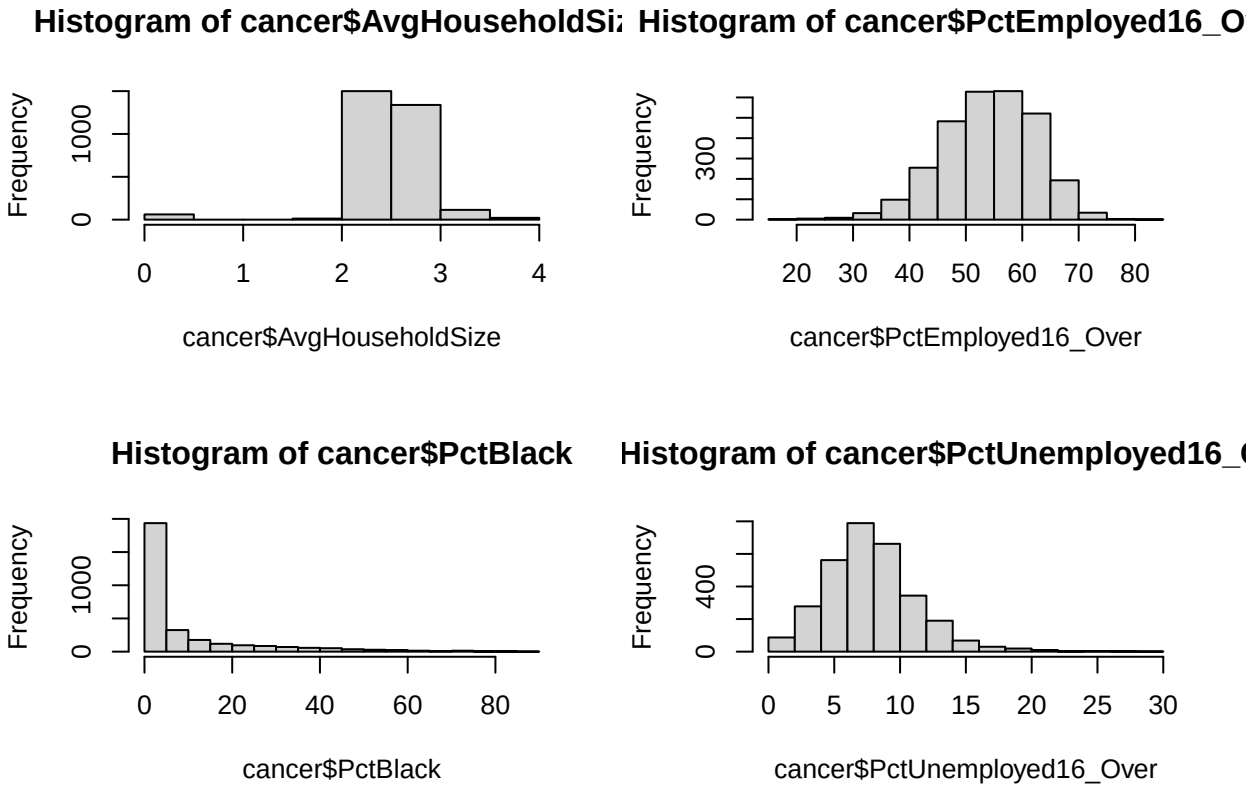
```
summary(cancer)
```

```
##   Geography         incidenceRate      medIncome                   binnedInc
##  Length:3047        Min.   : 201.3   Min.   : 22640   [22640, 34218.1]  : 306
##  Class :character   1st Qu.: 420.3   1st Qu.: 38883   (45201, 48021.6]  : 306
##  Mode  :character   Median : 453.5   Median : 45207   (54545.6, 61494.5]: 306
##                     Mean   : 448.3   Mean   : 47063   (42724.4, 45201]  : 305
##                     3rd Qu.: 480.9   3rd Qu.: 52492   (48021.6, 51046.4]: 305
##                     Max.   :1206.9   Max.   :125635   (51046.4, 54545.6]: 305
##                                                       (Other)           :1214
##  povertyPercent  MedianAgeMale   MedianAgeFemale AvgHouseholdSize
##  Min.   : 3.20   Min.   :22.40   Min.   :22.30   Min.   :0.0221
##  1st Qu.:12.15   1st Qu.:36.35   1st Qu.:39.10   1st Qu.:2.3700
##  Median :15.90   Median :39.60   Median :42.40   Median :2.5000
##  Mean   :16.88   Mean   :39.57   Mean   :42.15   Mean   :2.4797
##  3rd Qu.:20.40   3rd Qu.:42.50   3rd Qu.:45.30   3rd Qu.:2.6300
##  Max.   :47.40   Max.   :64.70   Max.   :65.70   Max.   :3.9700
##
##  PercentMarried  PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage
##  Min.   :23.10   Min.   :17.60      Min.   : 0.400       Min.   :22.30
##  1st Qu.:47.75   1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20
##  Median :52.40   Median :54.50      Median : 7.600       Median :65.10
##  Mean   :51.77   Mean   :54.15      Mean   : 7.852       Mean   :64.35
##  3rd Qu.:56.40   3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10
##  Max.   :72.50   Max.   :80.10      Max.   :29.400       Max.   :92.30
##                  NA's   :152
##  PctEmpPrivCoverage PctPublicCoverage    PctBlack       PctMarriedHouseholds
##  Min.   :13.5       Min.   :11.20     Min.   : 0.0000   Min.   :22.99
##  1st Qu.:34.5       1st Qu.:30.90     1st Qu.: 0.6207   1st Qu.:47.76
##  Median :41.1       Median :36.30     Median : 2.2476   Median :51.67
##  Mean   :41.2       Mean   :36.25     Mean   : 9.1080   Mean   :51.24
##  3rd Qu.:47.7       3rd Qu.:41.55     3rd Qu.:10.5097   3rd Qu.:55.40
##  Max.   :70.7       Max.   :65.10     Max.   :85.9478   Max.   :78.08
##
##     Edu18_24       deathRate
##  Min.   :1.487   Min.   : 59.7
##  1st Qu.:2.206   1st Qu.:161.2
##  Median :2.340   Median :178.1
##  Mean   :2.347   Mean   :178.7
##  3rd Qu.:2.486   3rd Qu.:195.2
##  Max.   :3.307   Max.   :362.8
##
```

We notice that there is one categorical variable (binnedInc) and one ID variable (Geography). Geography shouldn't be used as an explanatory variable but could be used as a visual aid.

We also note that PctEmployed16_Over has 152 missing values that should be investigated.

Univariate Plots:



We note that PctBlack has very large right skew and will need transforming to fix this.

We also note that there are some very suspicious low values in AvgHouseholdSize which are outliers. We create a new dataset containing these suspicious values.

```
sus <- subset(cancer, AvgHouseholdSize < 0.5)
head(sus)
```

```
## # A tibble: 6 x 18
##   Geography      incidenceRate medIncome binnedInc  povertyPercent MedianAgeMale
##   <chr>                  <dbl>     <dbl> <fct>                <dbl>         <dbl>
## 1 Berkeley Coun~           463.     56737 (54545.6,~            13.2          38
## 2 Grant County,~           356.     41039 (40362.7,~            17            45.5
## 3 Mineral Count~           454.     40714 (40362.7,~            19            48
## 4 Nye County, N~           454.     42881 (42724.4,~            16.8          50.8
## 5 Cayuga County~           471.     52792 (51046.4,~            12.7          41.2
## 6 Chenango Coun~           512.     46387 (45201, 4~            16.7          42.6
## # ... with 12 more variables: MedianAgeFemale <dbl>, AvgHouseholdSize <dbl>,
## #   PercentMarried <dbl>, PctEmployed16_Over <dbl>, PctUnemployed16_Over <dbl>,
## #   PctPrivateCoverage <dbl>, PctEmpPrivCoverage <dbl>,
## #   PctPublicCoverage <dbl>, PctBlack <dbl>, PctMarriedHouseholds <dbl>,
## #   Edu18_24 <dbl>, deathRate <dbl>
```

I will use an alternate source to check if the AvgHousehold Size is correct for Berkeley County.

```
sus[1,"AvgHouseholdSize"]
```

```
## # A tibble: 1 x 1
##   AvgHouseholdSize
##              <dbl>
## 1           0.0263
```

Reference: https://data.census.gov/cedsci/table?q=average%20household%20size&g=0500000US54003&y=2013&tid=ACSST1Y2013.S1101

From the same census data we have that the average household size of this county is 2.61. This indicates the data has been incorrectly inputted into this dataset or the data is missing. Either way, this incorrect data could cause the distribution to be unstable and should be considered as missing.

```
cancer1 <- cancer
susPoint <- which(cancer1$AvgHouseholdSize < 0.5)
cancer1$AvgHouseholdSize[susPoint] <- NA
summary(cancer1)
```

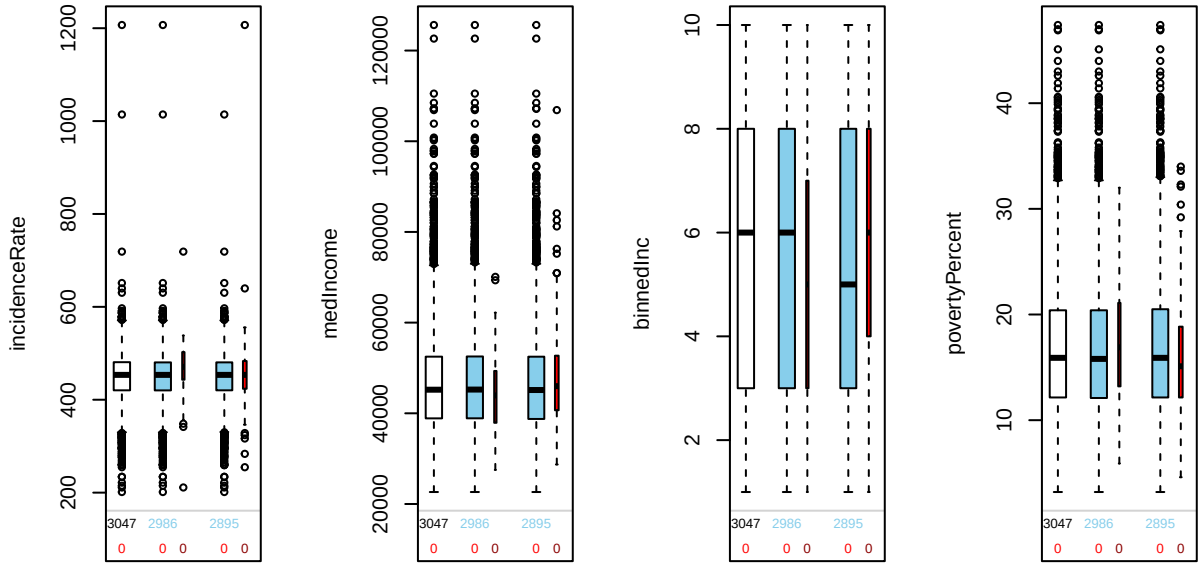
```
##   Geography         incidenceRate      medIncome                   binnedInc
##  Length:3047        Min.   : 201.3   Min.   : 22640   [22640, 34218.1]  : 306
##  Class :character   1st Qu.: 420.3   1st Qu.: 38883   (45201, 48021.6]  : 306
##  Mode  :character   Median : 453.5   Median : 45207   (54545.6, 61494.5]: 306
##                     Mean   : 448.3   Mean   : 47063   (42724.4, 45201]  : 305
##                     3rd Qu.: 480.9   3rd Qu.: 52492   (48021.6, 51046.4]: 305
##                     Max.   :1206.9   Max.   :125635   (51046.4, 54545.6]: 305
##                                                       (Other)           :1214
##  povertyPercent  MedianAgeMale   MedianAgeFemale AvgHouseholdSize
##  Min.   : 3.20   Min.   :22.40   Min.   :22.30   Min.   :1.86
##  1st Qu.:12.15   1st Qu.:36.35   1st Qu.:39.10   1st Qu.:2.37
##  Median :15.90   Median :39.60   Median :42.40   Median :2.50
##  Mean   :16.88   Mean   :39.57   Mean   :42.15   Mean   :2.53
##  3rd Qu.:20.40   3rd Qu.:42.50   3rd Qu.:45.30   3rd Qu.:2.64
##  Max.   :47.40   Max.   :64.70   Max.   :65.70   Max.   :3.97
##                                                  NA's   :61
##  PercentMarried  PctEmployed16_Over PctUnemployed16_Over PctPrivateCoverage
##  Min.   :23.10   Min.   :17.60      Min.   : 0.400       Min.   :22.30
##  1st Qu.:47.75   1st Qu.:48.60      1st Qu.: 5.500       1st Qu.:57.20
##  Median :52.40   Median :54.50      Median : 7.600       Median :65.10
##  Mean   :51.77   Mean   :54.15      Mean   : 7.852       Mean   :64.35
##  3rd Qu.:56.40   3rd Qu.:60.30      3rd Qu.: 9.700       3rd Qu.:72.10
##  Max.   :72.50   Max.   :80.10      Max.   :29.400       Max.   :92.30
##                  NA's   :152
##  PctEmpPrivCoverage PctPublicCoverage    PctBlack       PctMarriedHouseholds
##  Min.   :13.5       Min.   :11.20     Min.   : 0.0000   Min.   :22.99
##  1st Qu.:34.5       1st Qu.:30.90     1st Qu.: 0.6207   1st Qu.:47.76
##  Median :41.1       Median :36.30     Median : 2.2476   Median :51.67
##  Mean   :41.2       Mean   :36.25     Mean   : 9.1080   Mean   :51.24
##  3rd Qu.:47.7       3rd Qu.:41.55     3rd Qu.:10.5097   3rd Qu.:55.40
##  Max.   :70.7       Max.   :65.10     Max.   :85.9478   Max.   :78.08
##
##     Edu18_24       deathRate
##  Min.   :1.487   Min.   : 59.7
```

```
##  1st Qu.:2.206   1st Qu.:161.2
##  Median :2.340   Median :178.1
##  Mean   :2.347   Mean   :178.7
##  3rd Qu.:2.486   3rd Qu.:195.2
##  Max.   :3.307   Max.   :362.8
##
```

Missing Values and Strange Values:

There are now two predictor variables with what we would consider missing or strange values. We have to decide what kind of missing values these are. We can use plots from the VIM package to do this.

```
par(mfrow = c(1,4))
for (i in c(2:5)){
  pbox(cancer1,pos=i)
}
```

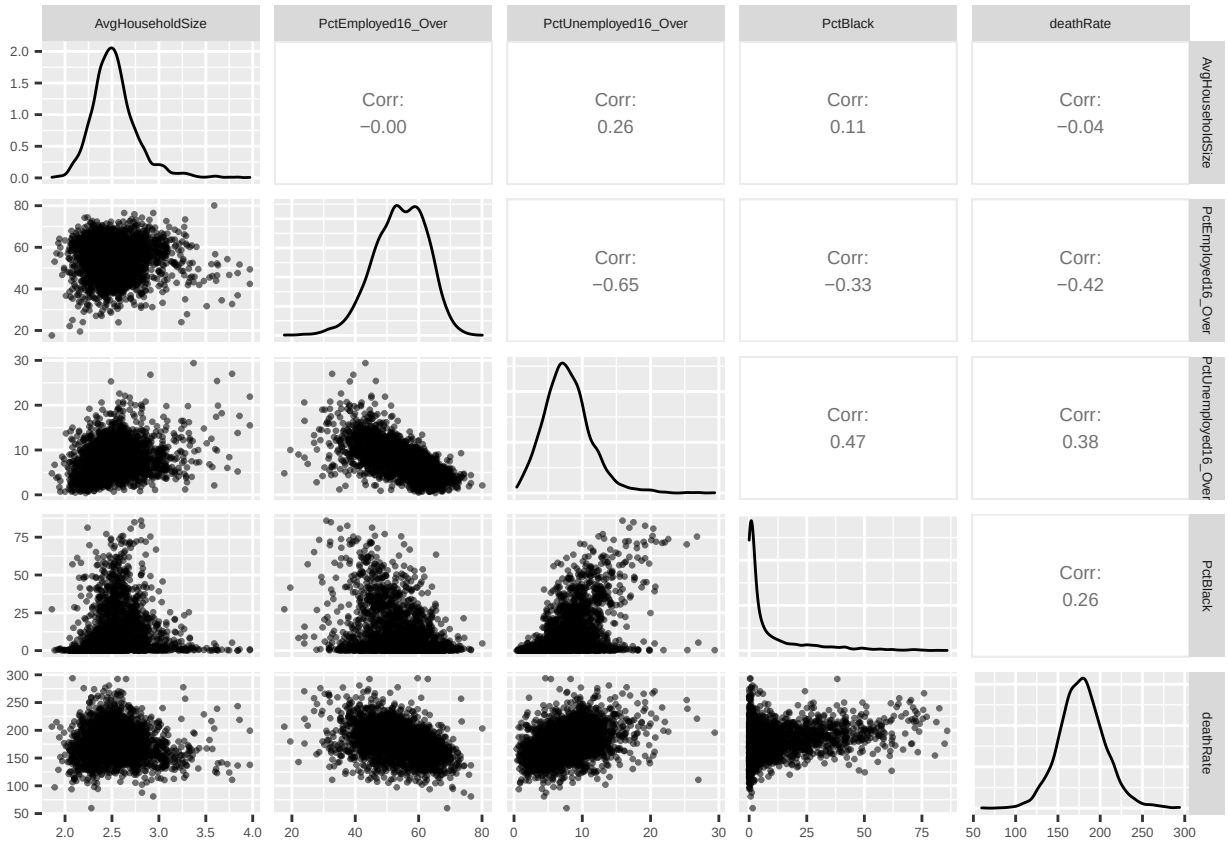


From the plots the difference in boxplots between missing values and non missing is quite similar across the board. We can consider these missing values MCAR and should just omit those rows with missing values as there isn't that many so should not cause our data to be unstable.

```
cancer2 <- na.omit(cancer1)
```

Comparing Predictors and Response

```
ggpairs(cancer2, columns = c(8,10,11,15,18), mapping = aes(alpha = 0.6),
        upper = list(continuous=wrap("cor", size=2.5, digits=2, stars=FALSE)),
        lower = list(continuous=wrap("points", size = 0.5)),
        diag = list(continuous="densityDiag"),
        progress = FALSE) +
  theme(text=element_text(size=6))
```



Unemployed and Employed 16 or over have high correlation with each other.

Below we have some plots of our variables against death rate:

PctBlack is incredibly right skewed, we can try a log transformation to fix it.

```
cancer2$blackFix <- log(cancer2$PctBlack)
ggplot(cancer2, aes(blackFix)) + geom_density()
```

```
## Warning: Removed 68 rows containing non-finite values (stat_density).
```

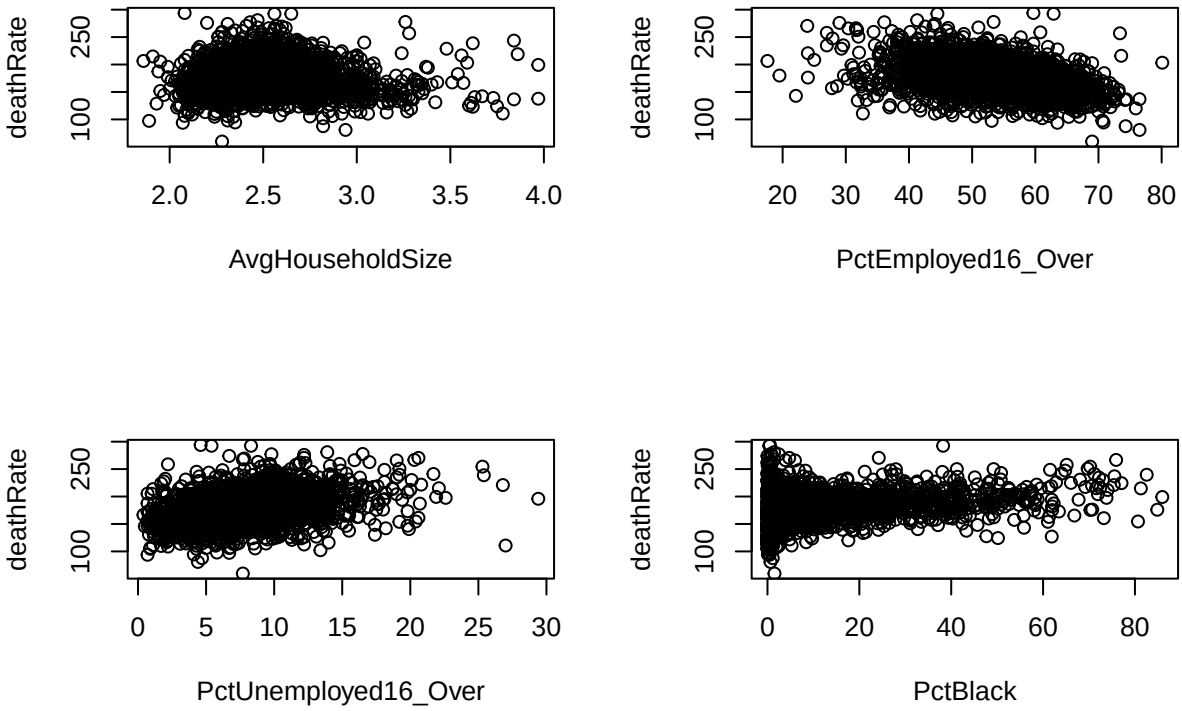
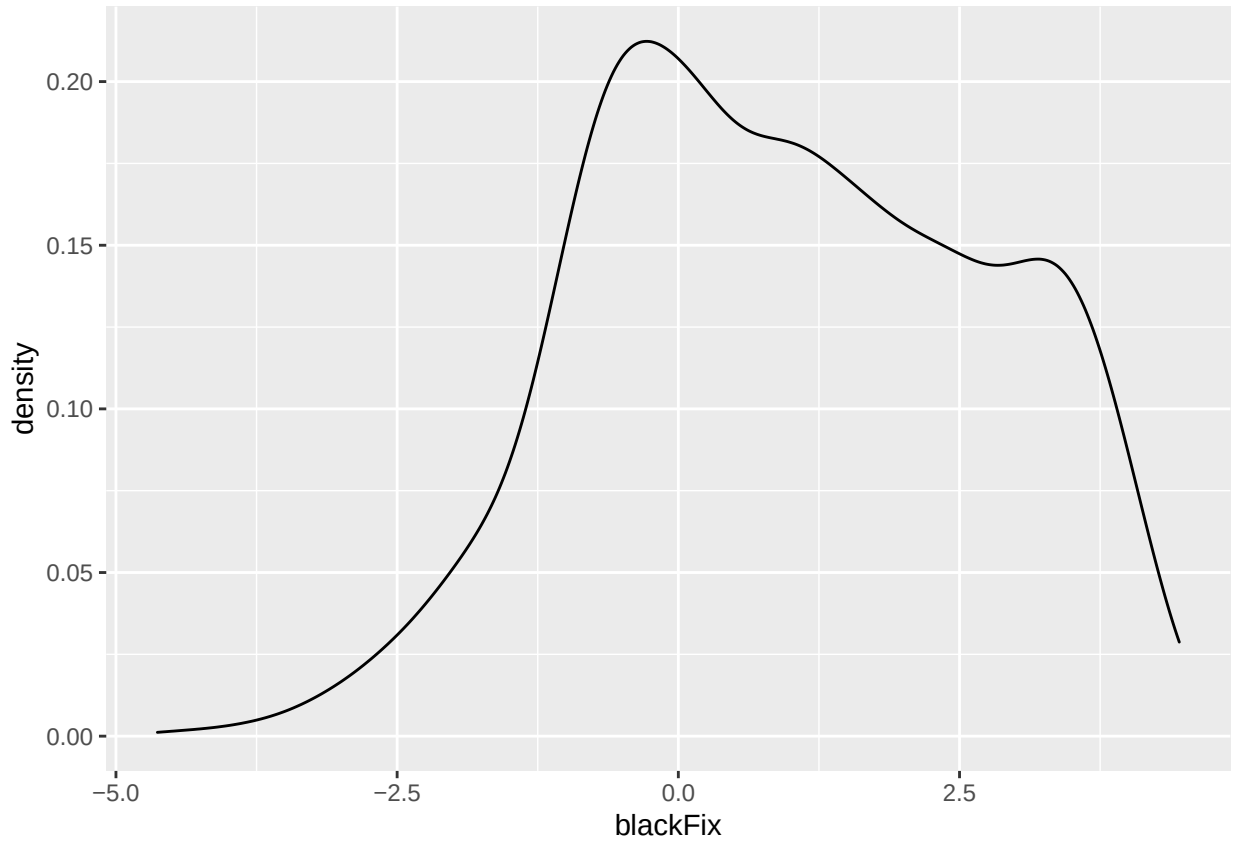
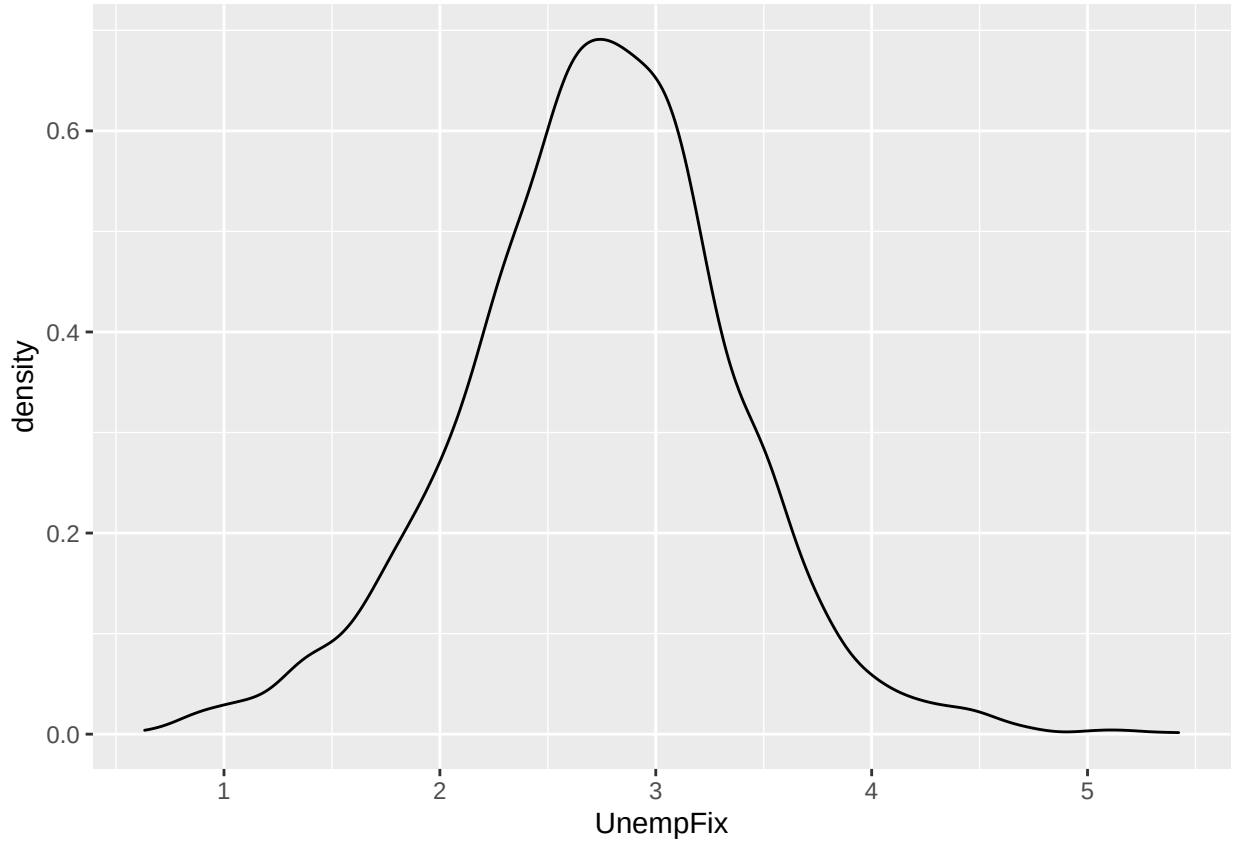


Figure 1: Predictors vs Death Plots



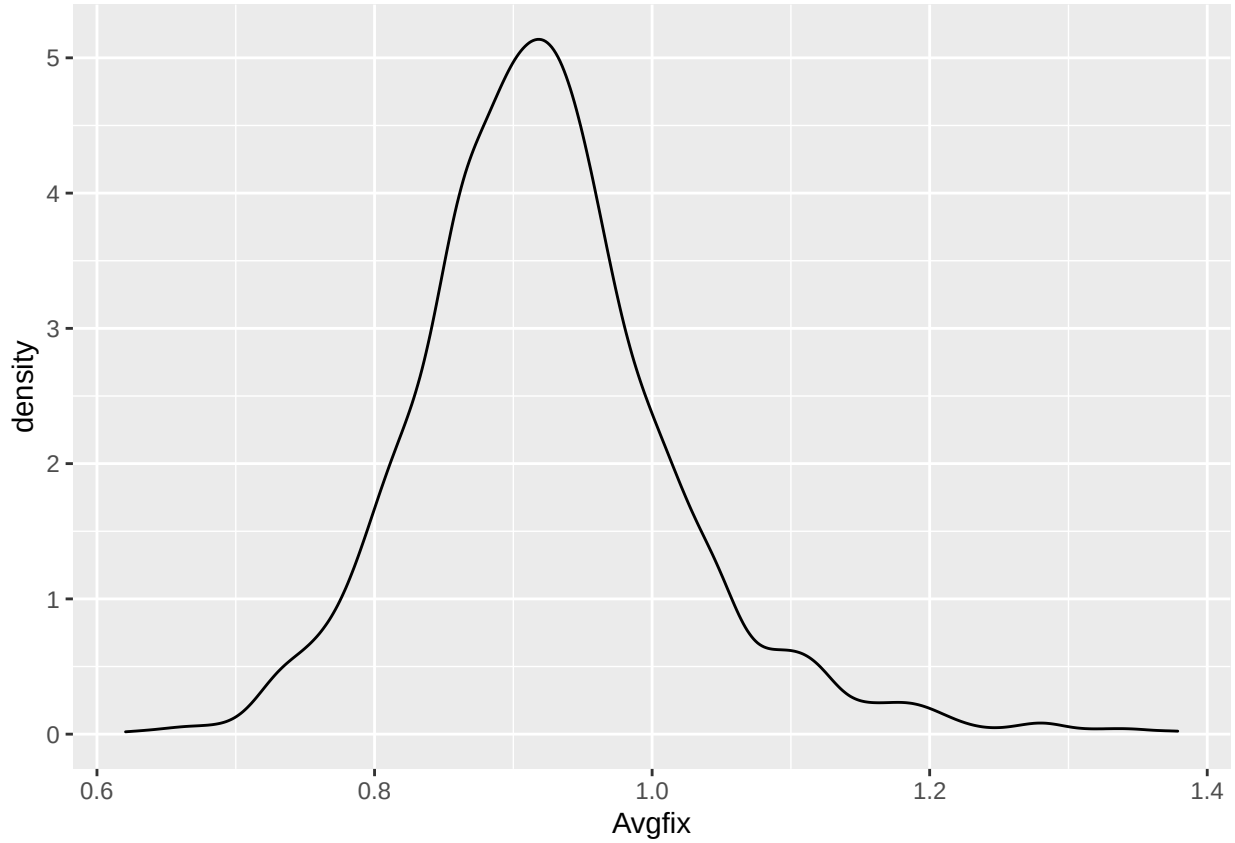
PctUnemployed16_Over is also slightly right skewed. We can try a square root transformation to fix this.

```
cancer2$UnempFix <- sqrt(cancer2$PctUnemployed16_Over)
ggplot(cancer2, aes(UnempFix)) + geom_density()
```



Average household size maybe needs fixing

```
cancer2$Avgfix <- log(cancer2$AvgHouseholdSize)
ggplot(cancer2, aes(Avgfix)) + geom_density()
```



We have evidence of multicolinearity between Unemployment and Employment.

We have evidence of a lot of heteroscedasticity for PctBlack, we can use a spread level plot for a suggested transformation

```
spreadLevelPlot(lm(deathRate~PctBlack, data = cancer2))
```

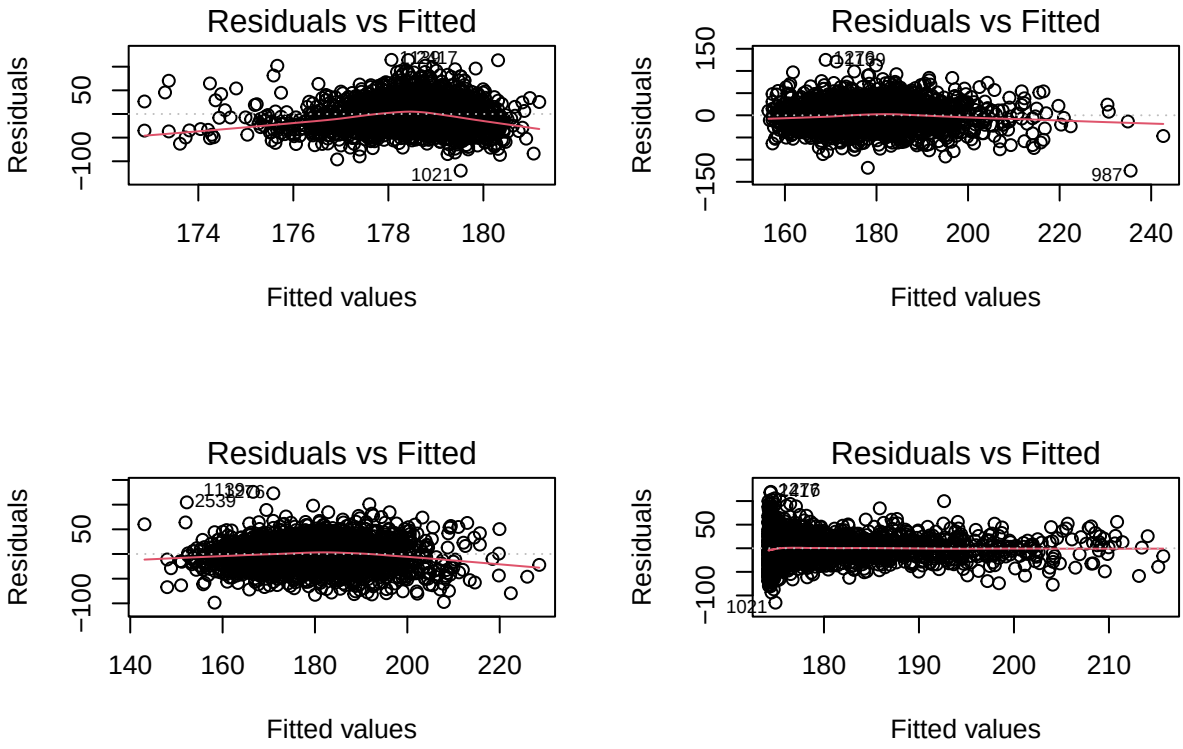
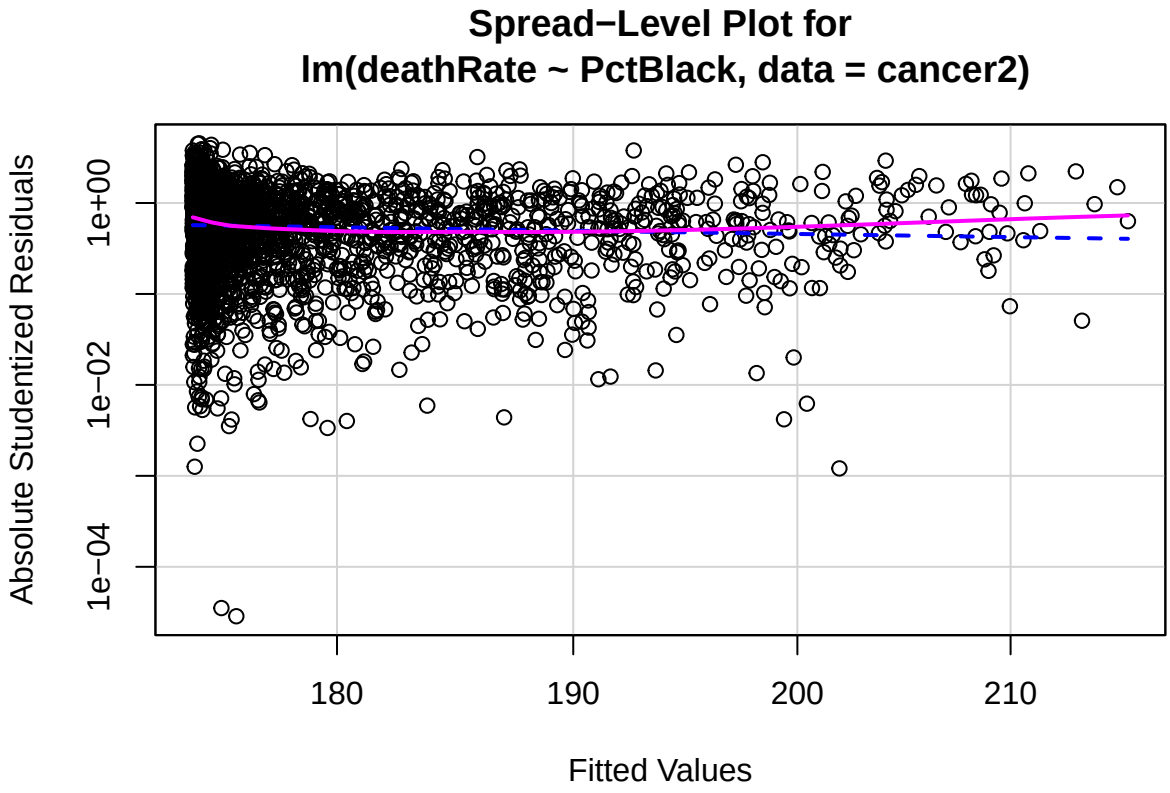


Figure 2: Residual Plots



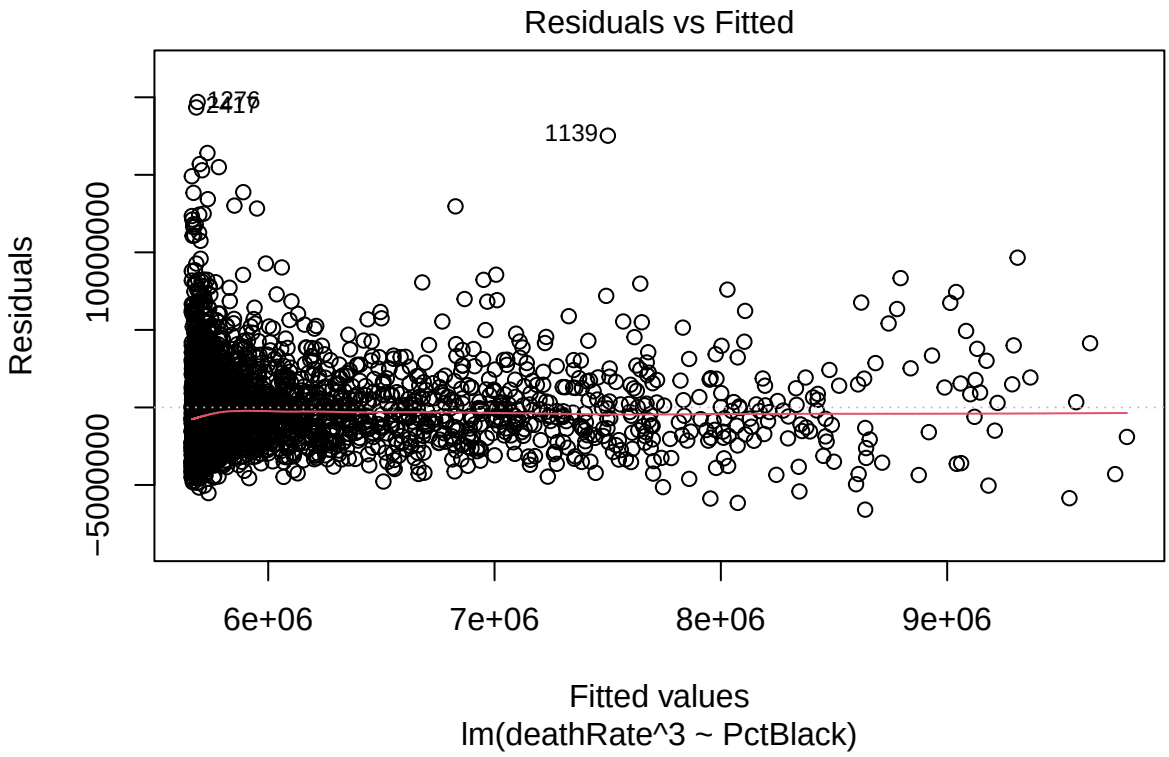

```
##
## Suggested power transformation:  2.616121
```

```
ncvTest(lm(deathRate~PctBlack, data = cancer2))
```

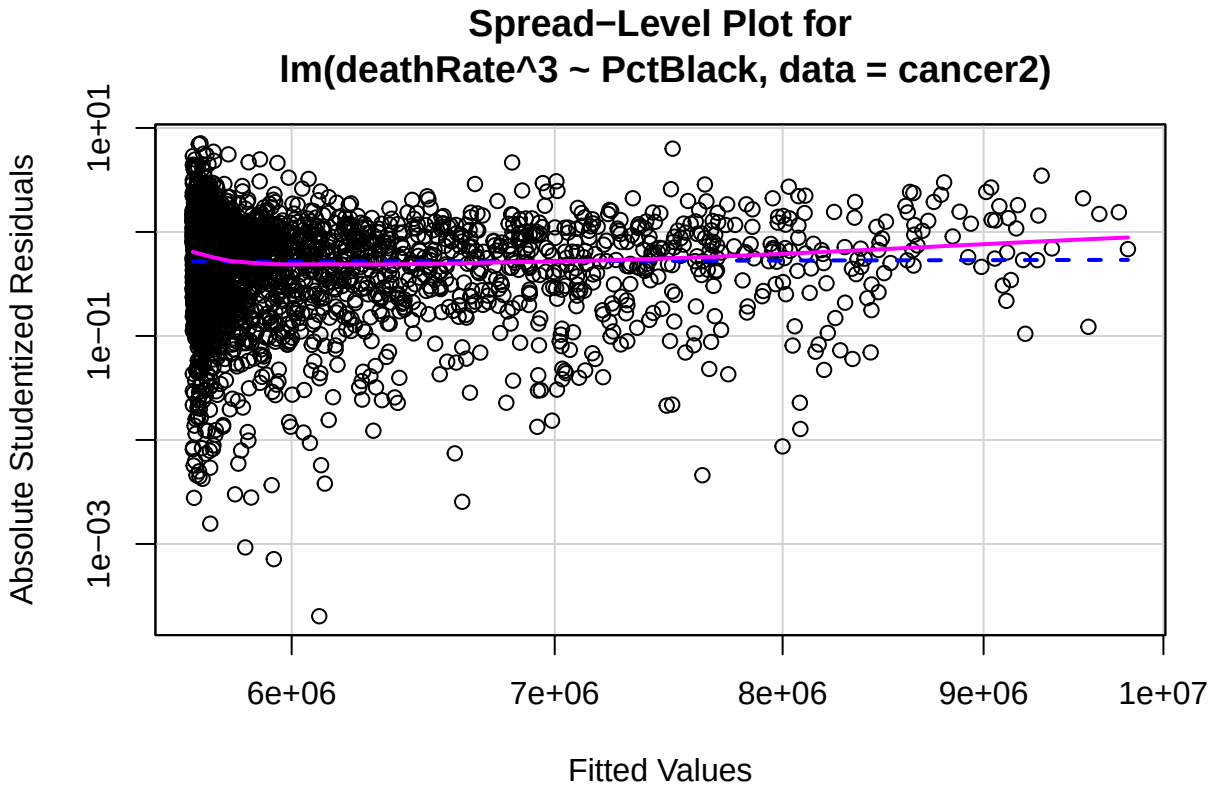
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 17.41301, Df = 1, p = 3.0076e-05
```

We try cubing the response to fix the heteroscedasticity

```
plot(lm(deathRate^3~PctBlack, data = cancer2), 1)
```




```
spreadLevelPlot(lm(deathRate^3~PctBlack, data = cancer2))
```




```
##
## Suggested power transformation:  0.9261115
```

This seems to have mostly fixed heteroscedasticity

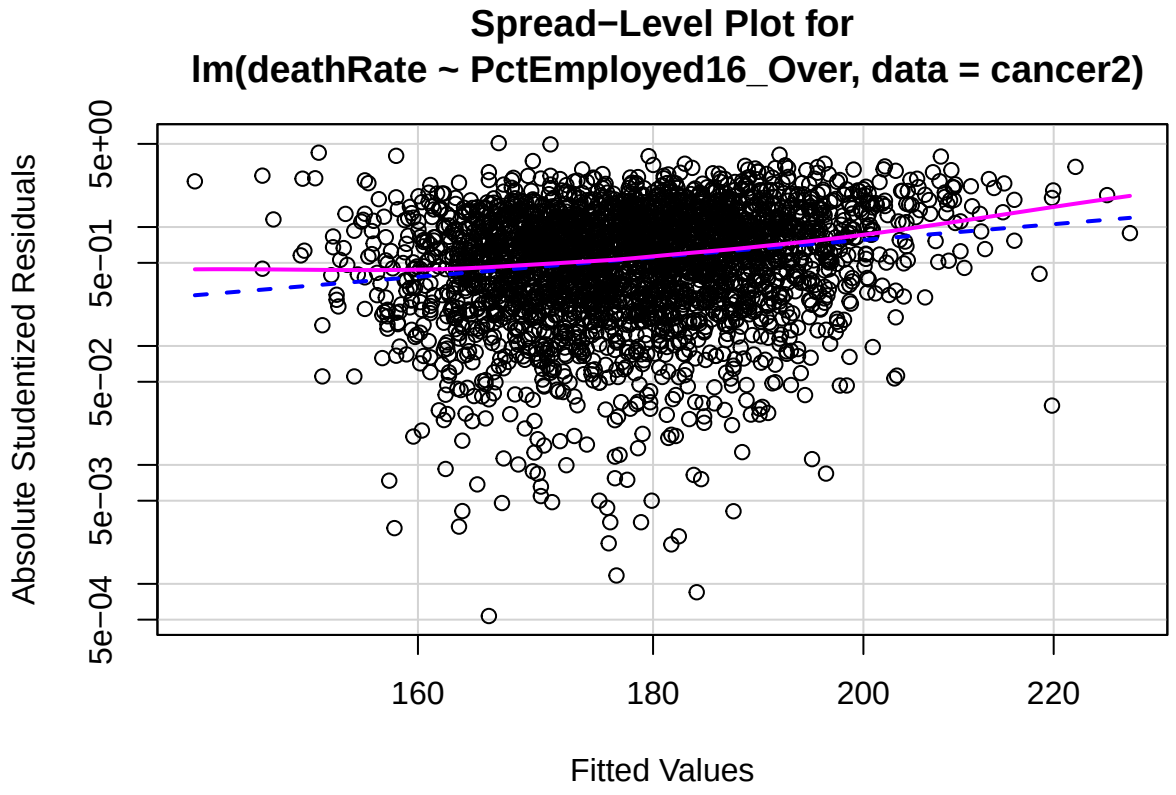
This is a different transformation than the one used to fix the density though

There is also some heteroscedasticity present for employed 16 or over.

```
ncvTest(lm(deathRate~PctEmployed16_Over, data = cancer2))
```

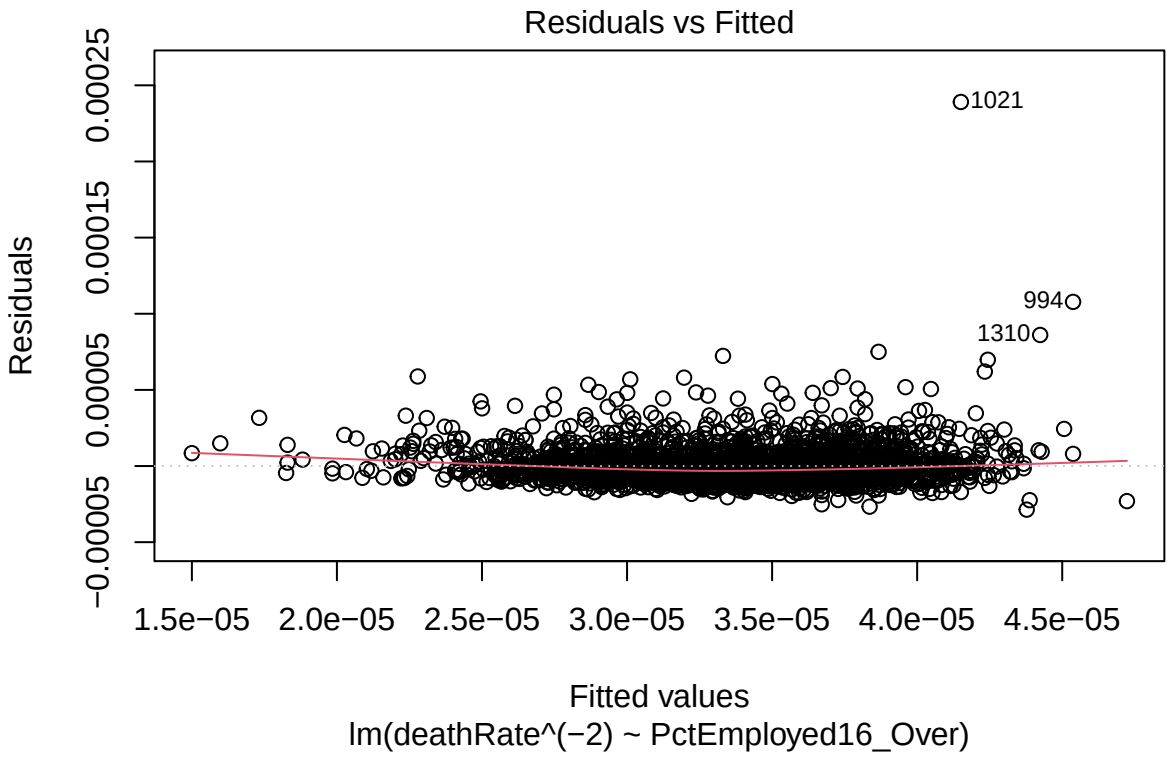
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 134.2628, Df = 1, p = < 2.22e-16
```

```
spreadLevelPlot(lm(deathRate~PctEmployed16_Over, data = cancer2))
```

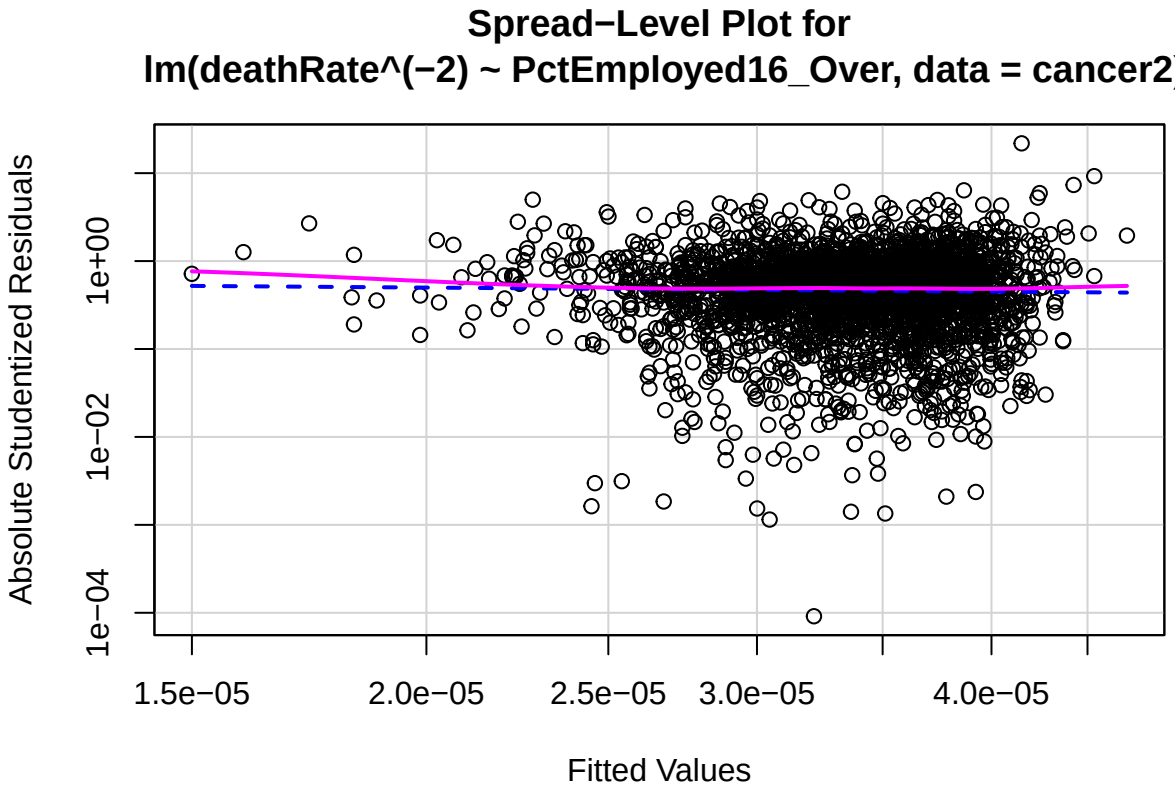



```
##
## Suggested power transformation:  -2.198456
```

```
plot(lm(deathRate^(-2)~PctEmployed16_Over, data = cancer2), 1)
```




```
spreadLevelPlot(lm(deathRate^(-2)~PctEmployed16_Over, data = cancer2))
```




```
##
## Suggested power transformation:  1.152844
```

Multicolinearity

We have high correlation between employed 16 and over and unemployed 16 and over. We will test to see if both are necessary.

```
vif(lm(deathRate~PctEmployed16_Over + PctUnemployed16_Over, data = cancer2))
```

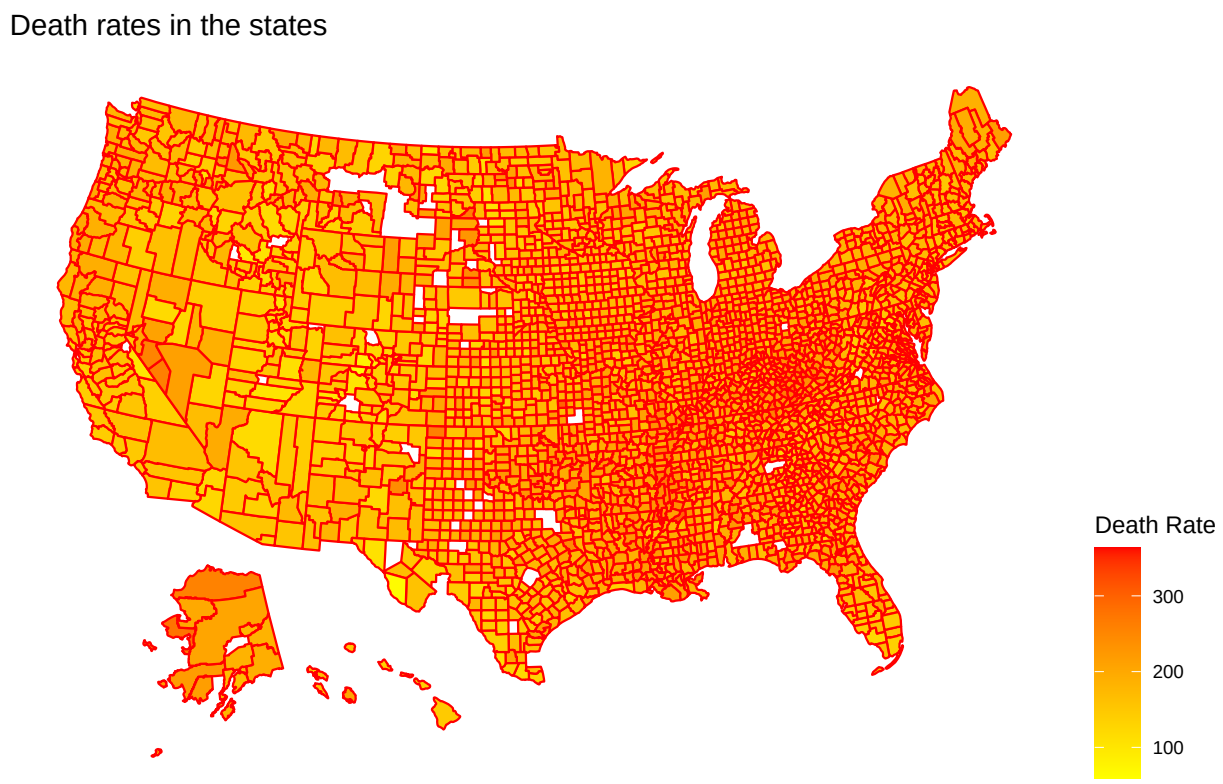
```
##   PctEmployed16_Over PctUnemployed16_Over
##             1.722034             1.722034
```

Variance inflation factors are low though, so we could potentially use both as they aren't measuring the exact same thing.

```
cancer3 <- cancer
uwu2 <- str_split(cancer3$Geography, pattern = ", ")
cancer3$state <- rep(0, length(cancer3$Geography))
for(i in 1:length(cancer3$Geography)){
  cancer3$state[i] <- uwu2[[i]][2]
}

cancer3$county <- rep(0, length(cancer3$Geography))
```

```
for(i in 1:length(cancer3$Geography)){
  cancer3$county[i] <- uwu2[[i]][1]
}
cancer3$county[167] <- "Dona Ana County"
cancer3$county[821] <- "La Salle Parish"
cancer3$fips <- rep("0", length(cancer3$Geography))
for(i in 1:length(cancer3$Geography)){
  cancer3$fips[i] <- fips(cancer3$state[i], cancer3$county[i])
}
deathMap <- subset(cancer3, select = c("fips", "county", "deathRate"))
plot_usmap(data = deathMap,regions = "counties", values = "deathRate", include = cancer3$fips, color =
  scale_fill_continuous(low = "yellow", high = "red", name = "Death Rate", label = scales::comma) +
  labs(title = "Death rates in the states") +
  theme(legend.position = "right")
```



The map above gives a nice illustration of what death rate looks like accross the United States, whites indicate that county was missing from the original dataset.

My recommendations for fixing would be apply the above transforms to fix skew, also remove the suspect and missing data as well to make sure our data isn't unstable.