# ST404 A1 Remus 1st

Remus

2022/2/4

### ST404

We are looking at 4 variables: povertyPercent, PctPrivateCoverage, PctEmpCoverage, PctPublicCoverage.

### Loading the data

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
```

### Summary of the data

```
##  povertyPercent  PctPrivateCoverage PctEmpPrivCoverage PctPublicCoverage
##  Min.   : 3.20   Min.   :22.30      Min.   :13.5       Min.   :11.20
##  1st Qu.:12.15   1st Qu.:57.20      1st Qu.:34.5       1st Qu.:30.90
##  Median :15.90   Median :65.10      Median :41.1       Median :36.30
##  Mean   :16.88   Mean   :64.35      Mean   :41.2       Mean   :36.25
##  3rd Qu.:20.40   3rd Qu.:72.10      3rd Qu.:47.7       3rd Qu.:41.55
##  Max.   :47.40   Max.   :92.30      Max.   :70.7       Max.   :65.10
```

### Missing data

```
print("The sum of missing values in these 4 variables are:")
```

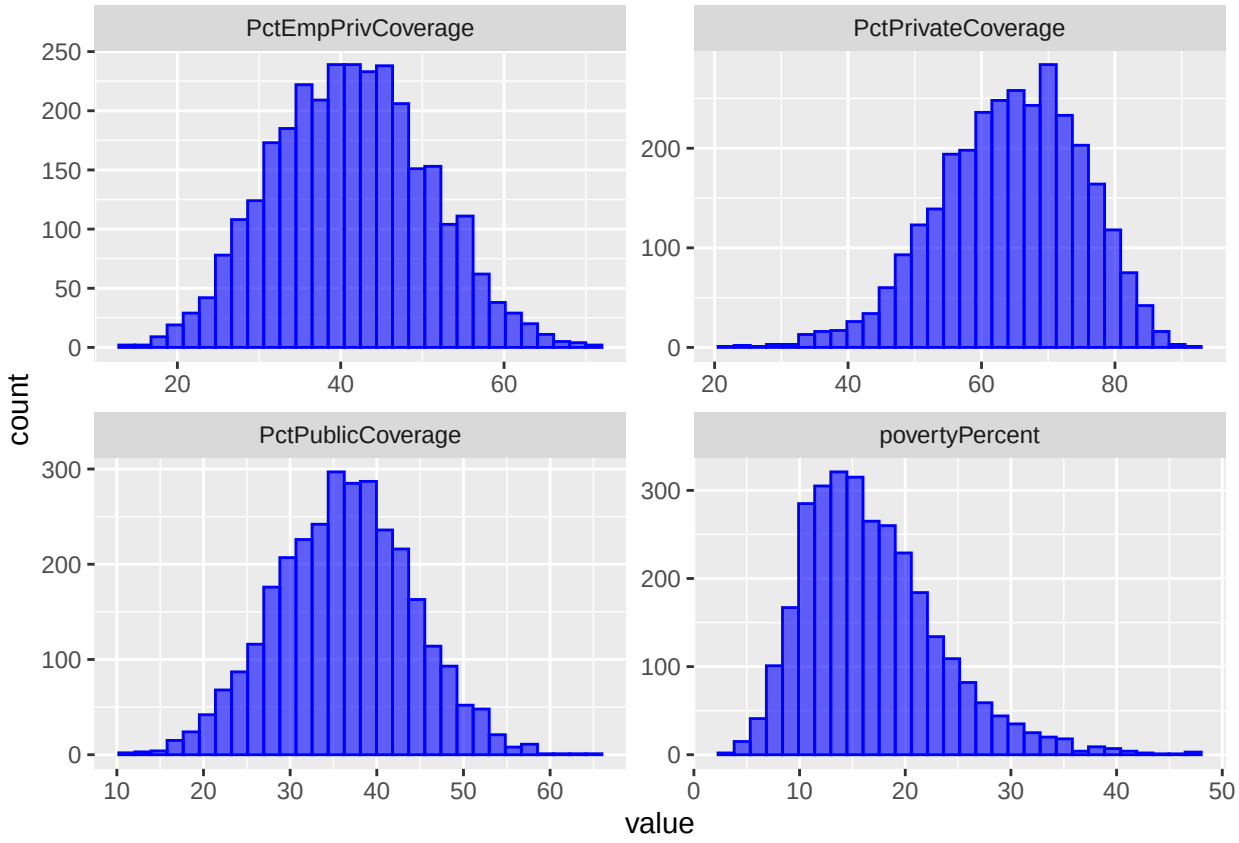
```
## [1] "The sum of missing values in these 4 variables are:"
```

```
sum(is.na(DF))
```

```
## [1] 0
```

## Univariate Plots

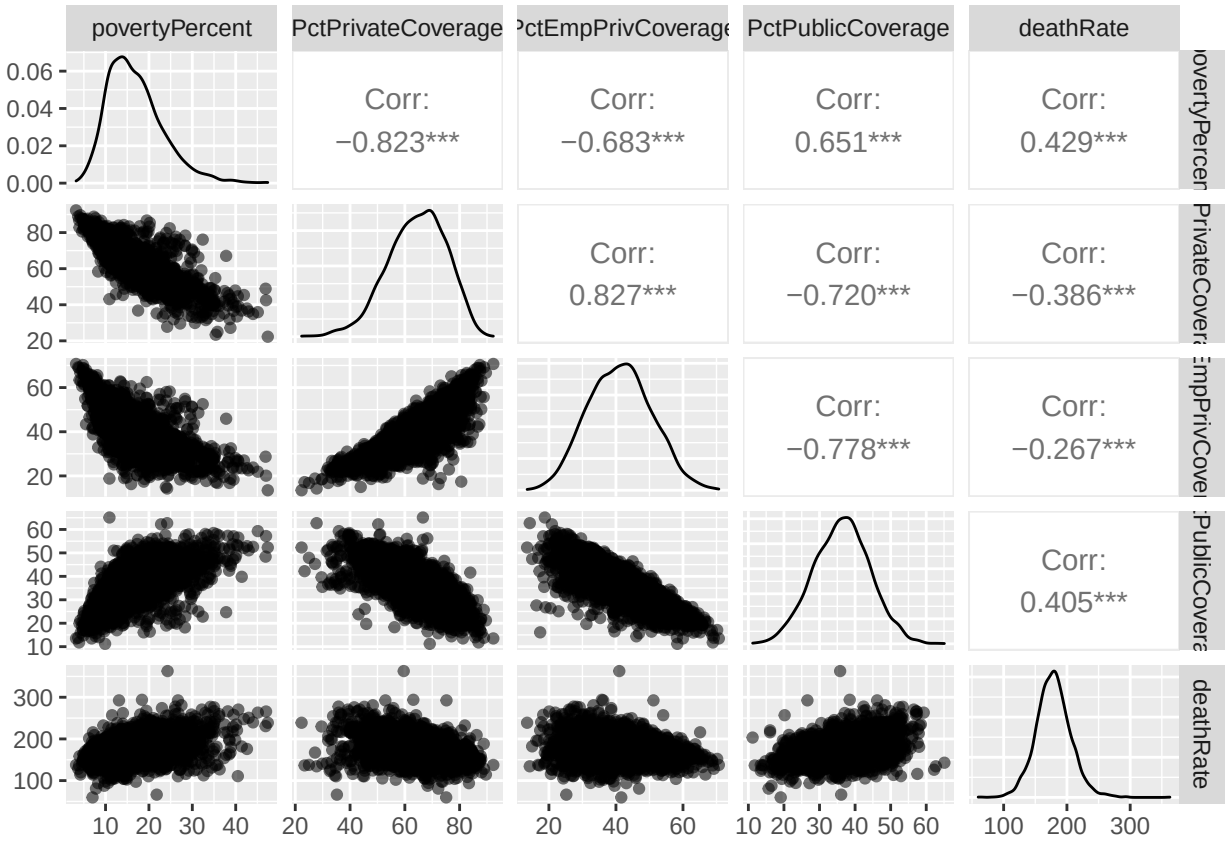
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




We can see that the three Coverage variables are symmetrical but povertyPercent is right-skewed. A transformation is needed to tackle this skewness.

## Predictors and Response

```
ggpairs(df[,c(5,12:14,18)], mapping  = aes(alpha = 0.3))
```

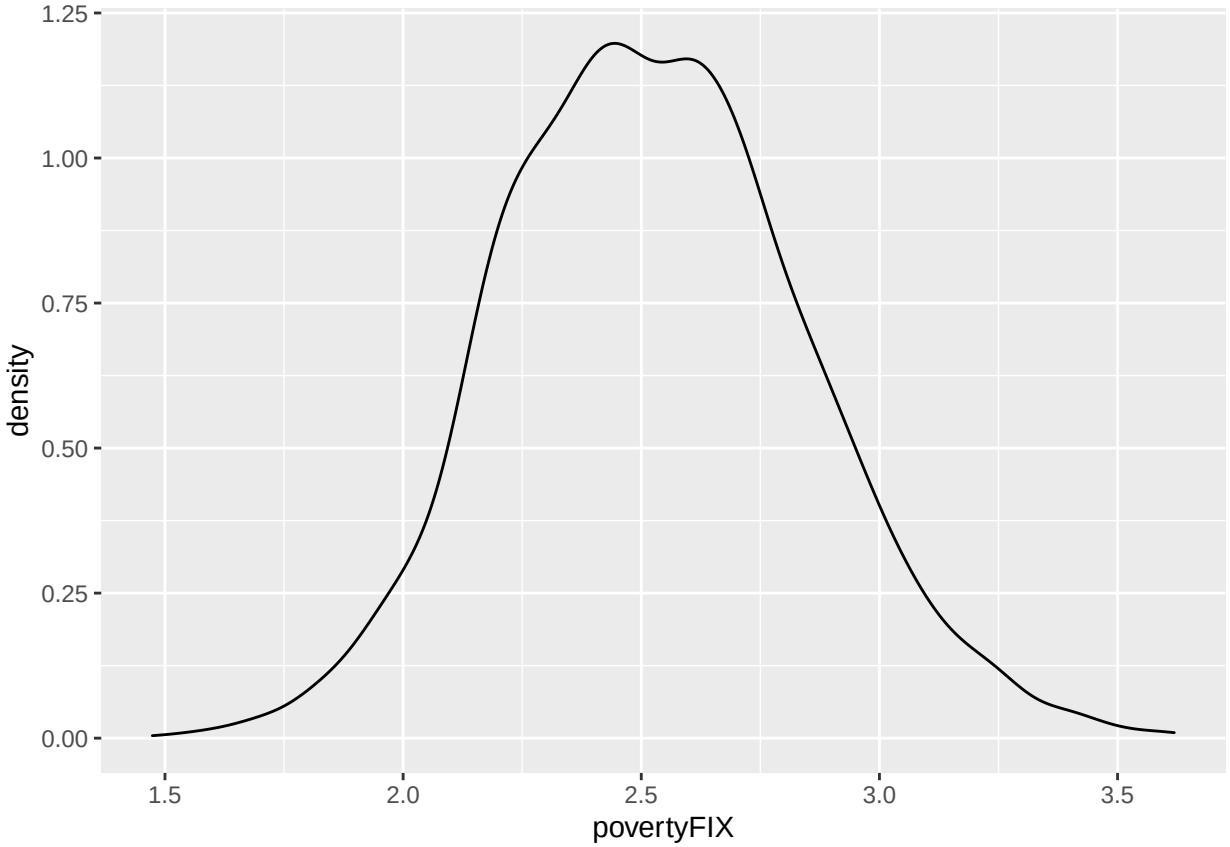



The three coverage variables are highly correlated as expected. However, the deathRate is not significantly correlated to any of the coverage variables. The correlation coefficicent between deathRate and the coverage variables do not exceed 0.45.

## Transformation

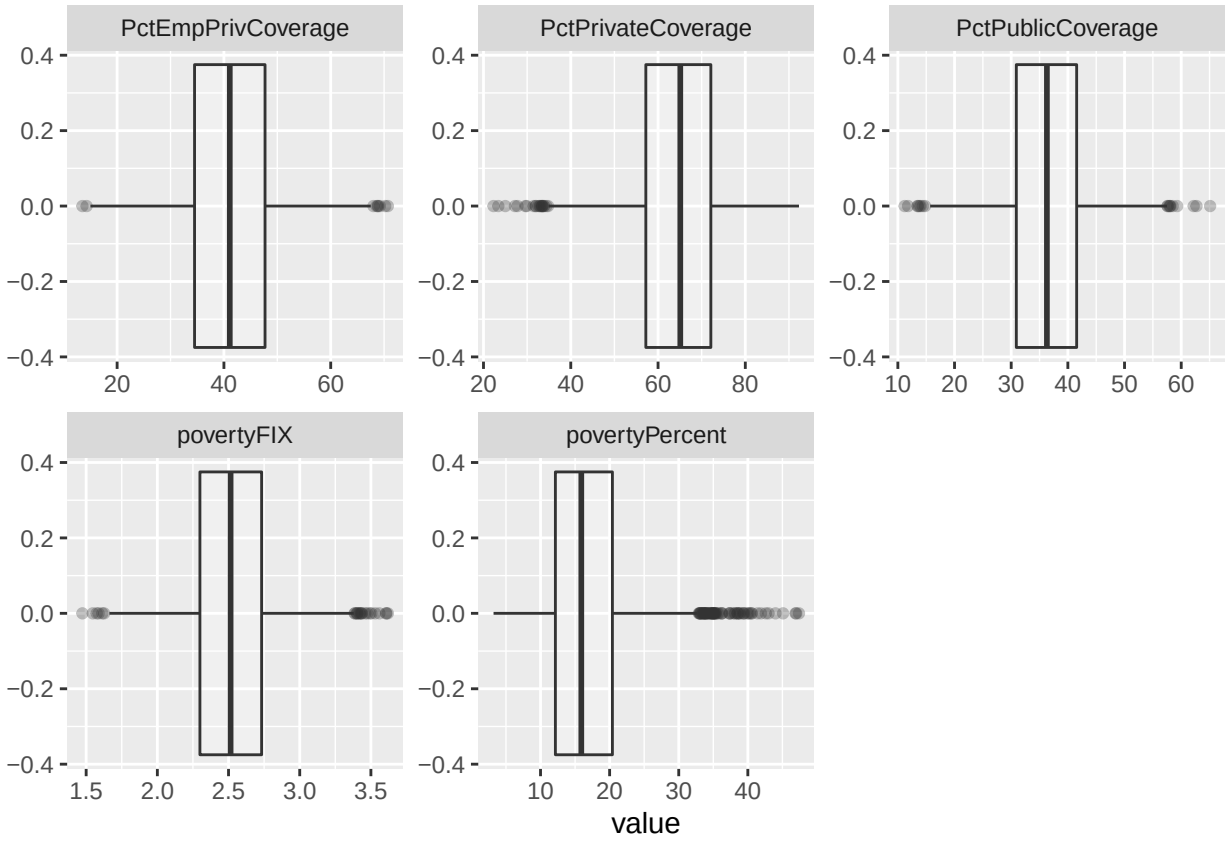
Try cube rooting.

```
DF$povertyFIX <- (DF$povertyPercent)^(1/3)
ggplot(DF, aes(povertyFIX))+geom_density()
```




## Outliers

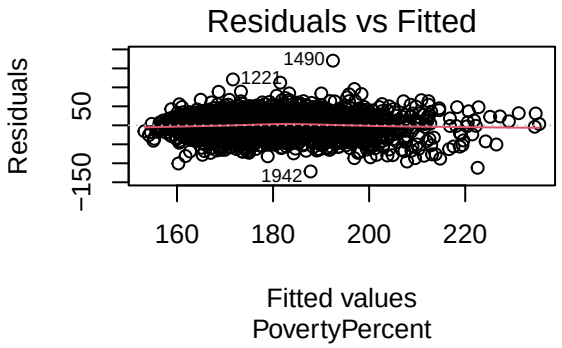
First we use box-plots to see if there are any suspicous value.

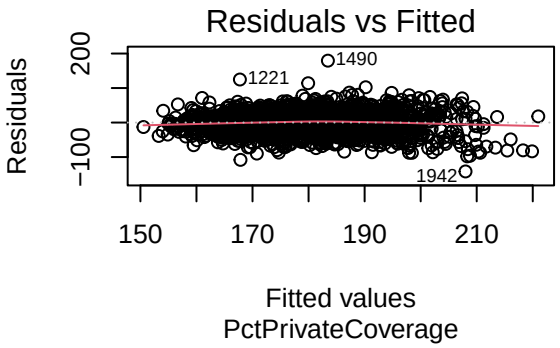


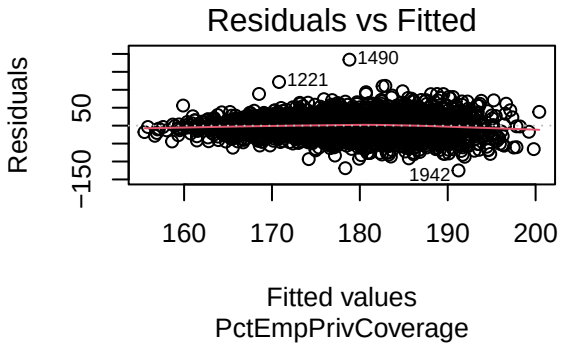

From the box-plots, all four variables have a quite a big number of extreme values. We need further investigations.

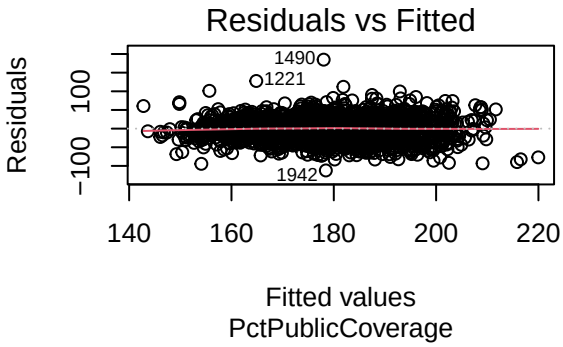
## Hetroscedascity

We can look at the residuals plots of these variables in simple models.



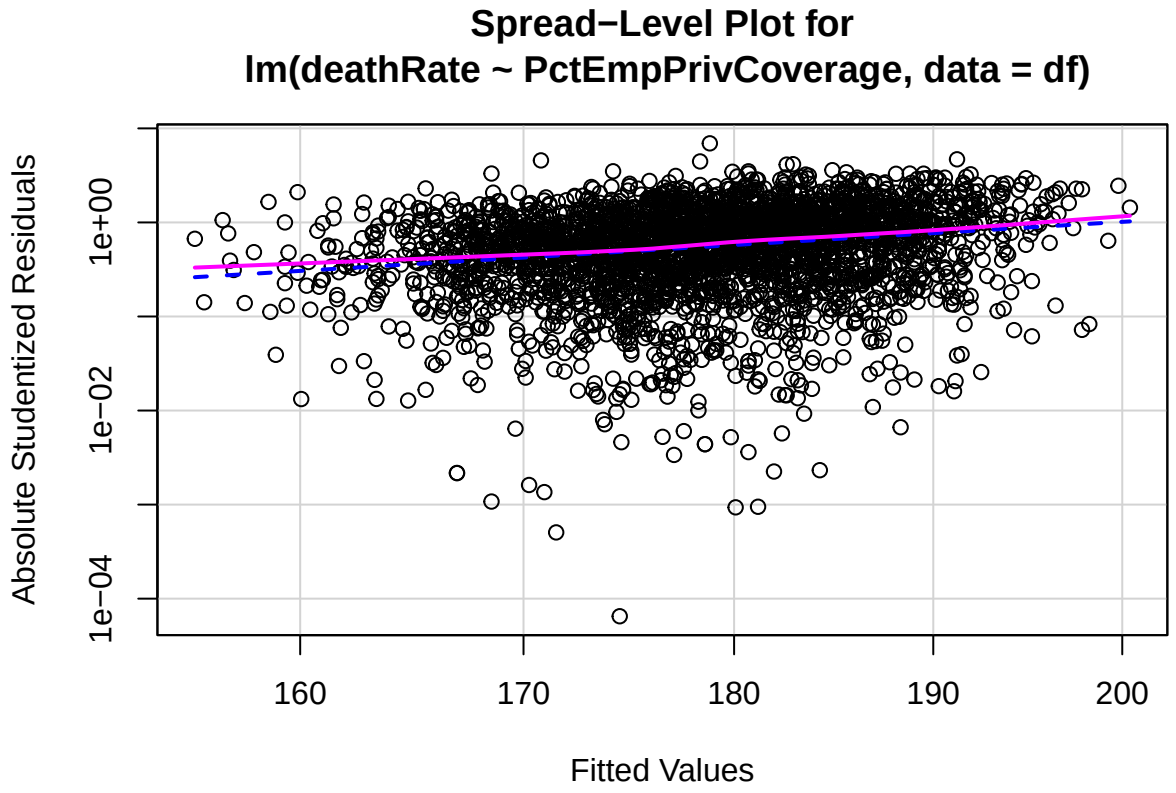




There is potential heteroscedascity in PctEmpPrivCoverage.

```
ncvTest(lm(deathRate~PctEmpPrivCoverage, data = cancer))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 228.6336, Df = 1, p = < 2.22e-16
```

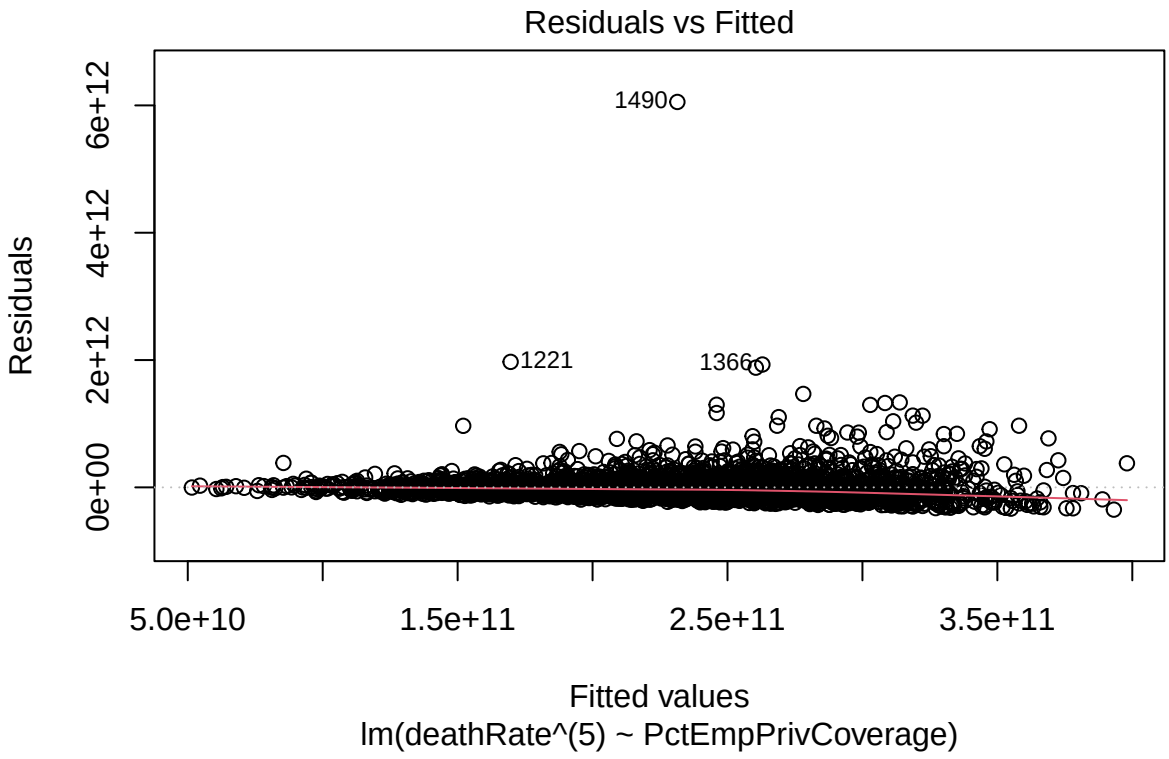
```
spreadLevelPlot(lm(deathRate~PctEmpPrivCoverage, data = df))
```




```
##
## Suggested power transformation:  -4.382922
```

Try power=5

```
plot(lm(deathRate^(5)~PctEmpPrivCoverage, data = df),1)
```




```
ncvTest(lm(deathRate^5~PctEmpPrivCoverage, data = df))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 361.1926, Df = 1, p = < 2.22e-16
```

```
vif(lm(deathRate~PctPrivateCoverage+PctEmpPrivCoverage+ PctPublicCoverage, data = df))
```

```
## PctPrivateCoverage PctEmpPrivCoverage  PctPublicCoverage
##           3.325961           4.062968           2.660187
```

```
vif(lm(deathRate~PctPrivateCoverage+PctEmpPrivCoverage+ PctPublicCoverage+ povertyPercent, data = df))
```

```
## PctPrivateCoverage PctEmpPrivCoverage  PctPublicCoverage     povertyPercent
##           5.255494           4.083639           2.734522           3.178377
```

Although VIF are not very big,the three healthcare coverage variables are clearly correlated. I would discard PctPrivateCoverage and PctEmpPrivCoverage in fitting a model. I leave PctPublicCoverage because it is the least correlated variable with povertyPercent among the healthcare coverage variables.