Linear regression predicting covid19 death rate

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.0.3
```

```r
library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:car':
##
##     logit
```

```r
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(nortest)
```

```
## Warning: package 'nortest' was built under R version 4.0.3
```

```r
mydata<-read.csv("flInfo.csv")

newdata<-read.csv("https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/asrh/cc-e
```

```r
library(sqldf)
```

```
## Loading required package: gsubfn
```

```
## Loading required package: proto
```

```
## Loading required package: RSQLite
```

```r
#Select the total population and population of age65+ for 2019(Year=10)
mydata2<-sqldf("SELECT CTYNAME, POPESTIMATE, AGE65PLUS_TOT, POPEST_MALE, POPEST_FEM
      FROM newdata
      WHERE YEAR=10")
```

```r
#calculate the old age rate for these countries
mydata2$ElderlyRate<-mydata2$AGE65PLUS_TOT/mydata2$POPESTIMATE
#calculate the gender rate for these countries(Male Population/Female Population*100%)
mydata2$GenderRate<-mydata2$POPEST_MALE/mydata2$POPEST_FEM
```

```r
#reformating the name of countries
mydata2$CITYYNAME = as.character(mydata2$CTYNAME)
mydata2$admin <- substr(mydata2$CTYNAME,0, nchar(mydata2$CITYYNAME)-7)
```

```r
#Join the two tables
mydata<-sqldf("SELECT *
              FROM mydata2,mydata
              WHERE mydata2.admin=mydata.Admin2")
```

```r
#remove the duplicate
mydata<-mydata[,-c(1,8)]
```

```r
unemploymentdata<-read.csv("https://raw.githubusercontent.com/dawnzyf/Florida-unemployment-dataset/main/
```

```r
#SELECT data of 2020 Aug
unemploymentdata<-unemploymentdata[1:67,1:5]
```

```r
#Formatting county name
unemploymentdata$COUNTY = as.character(unemploymentdata$COUNTY)
unemploymentdata$COUNTY<-substr(unemploymentdata$COUNTY,0,nchar(unemploymentdata$COUNTY)-7)
```

```r
#join into dataset
mydata<-sqldf("SELECT *
              FROM unemploymentdata,mydata
              WHERE mydata.Admin2=unemploymentdata.COUNTY")
```

```r
#View datatype
str(mydata)
```

```
## 'data.frame':    67 obs. of  32 variables:
##  $ COUNTY        : chr  "Alachua" "Baker" "Bay" "Bradford" ...
##  $ FORCE         : num  132853 11553 82171 10839 285860 ...
##  $ MENT          : num  126345 11047 77943 10296 268133 ...
##  $ LEVEL         : num  6508 506 4228 543 17727 ...
##  $ RATE          : num  0.049 0.044 0.051 0.05 0.062 0.092 0.048 0.067 0.073 0.046 ...
##  $ POPESTIMATE   : int  266309 28254 184736 27142 587769 1934516 14428 181522 145415 212228 ...
##  $ AGE65PLUS_TOT : int  36454 3899 31147 4854 136862 314039 2580 71802 52553 32388 ...
##  $ POPEST_MALE   : int  128758 14874 91977 14744 287247 942414 7870 88453 70393 104587 ...
##  $ POPEST_FEM    : int  137551 13380 92759 12398 300522 992102 6558 93069 75022 107641 ...
##  $ ElderlyRate   : num  0.137 0.138 0.169 0.179 0.233 ...
##  $ GenderRate    : num  0.936 1.112 0.992 1.189 0.956 ...
##  $ admin         : chr  "Alachua" "Baker" "Bay" "Bradford" ...
##  $ Index         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Admin2        : chr  "Alachua" "Baker" "Bay" "Bradford" ...
##  $ FIPS          : int  12001 12003 12005 12007 12009 12011 12013 12015 12017 12019 ...
##  $ Province_State: chr  "Florida" "Florida" "Florida" "Florida" ...
##  $ Country_Region: chr  "US" "US" "US" "US" ...
##  $ Last_Update   : chr  "10/2/2020 4:23" "10/2/2020 4:23" "10/2/2020 4:23" "10/2/2020 4:23" ...
##  $ Latitude      : num  29.7 30.3 30.3 30 28.3 ...
##  $ Longitude     : num  -82.4 -82.3 -85.6 -82.2 -80.7 ...
##  $ Confirmed     : int  8352 1492 6146 1135 9154 77433 622 3097 2793 5193 ...
##  $ Deaths        : int  61 13 111 9 299 1406 12 136 107 103 ...
##  $ Recovered     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Active        : int  8291 1479 6035 1126 8855 76027 610 2961 2686 5090 ...
##  $ Combined_Key  : chr  "Alachua, Florida, US" "Baker, Florida, US" "Bay, Florida, US" "Bradford, Fl
##  $ Incidence_Rate: num  3104 5108 3518 4025 1521 ...
##  $ lethality     : num  0.0073 0.00871 0.01806 0.00793 0.03266 ...
##  $ Population    : int  269043 29210 174705 28201 601942 1952778 14105 188910 149657 219252 ...
##  $ ConperCapita  : num  0.031 0.0511 0.0352 0.0402 0.0152 ...
##  $ DeathsPerCap  : num  0.000227 0.000445 0.000635 0.000319 0.000497 ...
##  $ SqMiles       : num  875 585 758 294 1016 ...
##  $ PopDensity    : num  307.5 49.9 230.3 95.9 592.7 ...
```

```r
#Remove duplicate
mydata<-mydata[,c(-12,-13)]
```

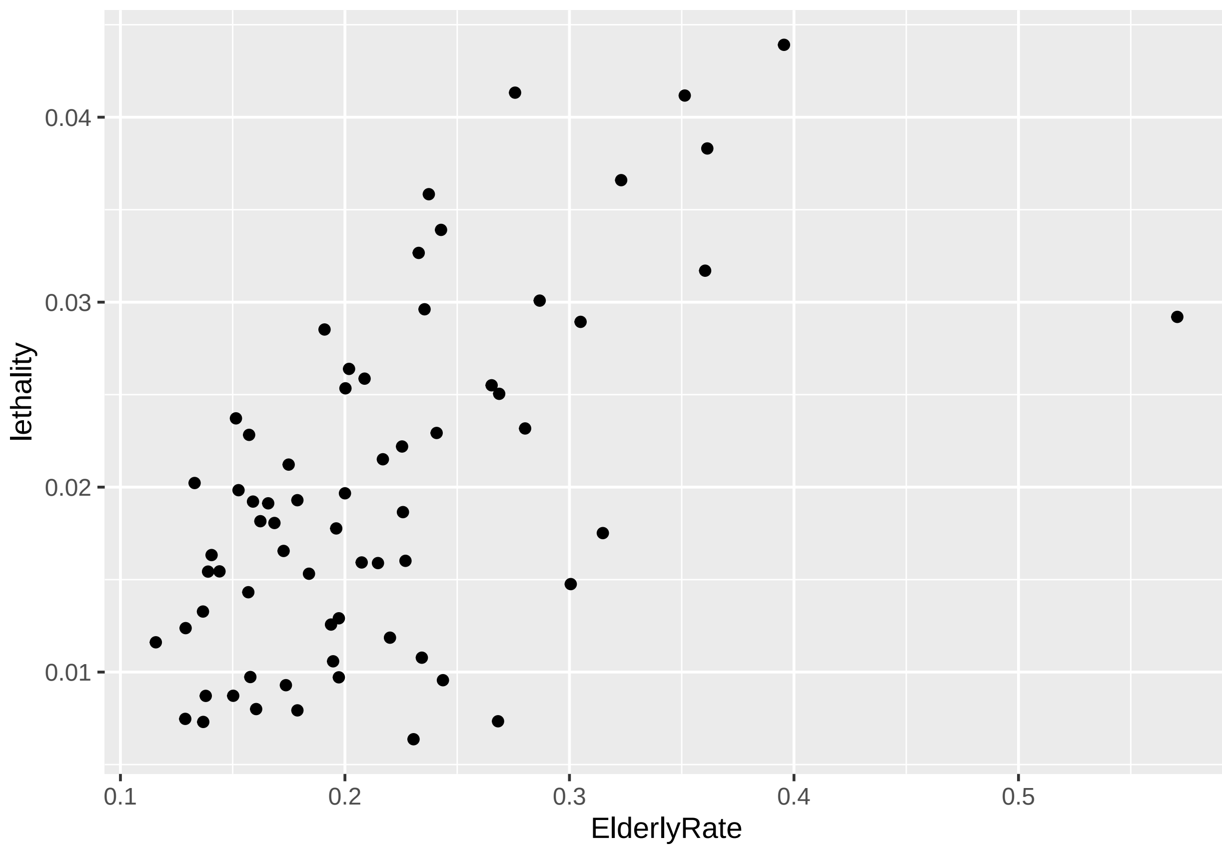
```r
#rename column
colnames(mydata)[c(3,4,5)]<-c("Labor","UnemploymentPOP","UnemployRate")
```

```r
#View data
head(mydata)
```

```
##     COUNTY   FORCE  Labor UnemploymentPOP UnemployRate POPESTIMATE
```

```
## 1  Alachua  132853 126345           6508        0.049      266309
## 2    Baker   11553  11047            506        0.044       28254
## 3      Bay   82171  77943           4228        0.051      184736
## 4 Bradford   10839  10296            543        0.050       27142
## 5  Brevard  285860 268133          17727        0.062      587769
## 6  Broward 1015939 922021          93918        0.092     1934516
##   AGE65PLUS_TOT POPEST_MALE POPEST_FEM ElderlyRate GenderRate   Admin2  FIPS
## 1         36454      128758     137551   0.1368861  0.9360746  Alachua 12001
## 2          3899       14874      13380   0.1379982  1.1116592    Baker 12003
## 3         31147       91977      92759   0.1686028  0.9915696      Bay 12005
## 4          4854       14744      12398   0.1788372  1.1892241 Bradford 12007
## 5        136862      287247     300522   0.2328500  0.9558269  Brevard 12009
## 6        314039      942414     992102   0.1623347  0.9499164  Broward 12011
##   Province_State Country_Region    Last_Update Latitude Longitude Confirmed
## 1        Florida             US 10/2/2020 4:23 29.67867 -82.35928      8352
## 2        Florida             US 10/2/2020 4:23 30.33060 -82.28467      1492
## 3        Florida             US 10/2/2020 4:23 30.26549 -85.62123      6146
## 4        Florida             US 10/2/2020 4:23 29.95080 -82.16612      1135
## 5        Florida             US 10/2/2020 4:23 28.29410 -80.73091      9154
## 6        Florida             US 10/2/2020 4:23 26.15185 -80.48726     77433
##   Deaths Recovered Active          Combined_Key Incidence_Rate   lethality
## 1     61         0   8291  Alachua, Florida, US       3104.336 0.007303640
## 2     13         0   1479    Baker, Florida, US       5107.840 0.008713137
## 3    111         0   6035      Bay, Florida, US       3517.930 0.018060527
## 4      9         0   1126 Bradford, Florida, US       4024.680 0.007929515
## 5    299         0   8855  Brevard, Florida, US       1520.745 0.032663317
## 6   1406         0  76027  Broward, Florida, US       3965.274 0.018157633
##   Population ConperCapita DeathsPerCap SqMiles PopDensity
## 1     269043   0.03104336  0.000226730  875.02  307.47069
## 2      29210   0.05107840  0.000445053  585.23   49.91200
## 3     174705   0.03517930  0.000635357  758.46  230.34175
## 4      28201   0.04024680  0.000319138  293.96   95.93482
## 5     601942   0.01520744  0.000496726 1015.66  592.66093
## 6    1952778   0.03965274  0.000720000 1209.78 1614.15960
```

```
library(ggplot2)
ggplot(mydata)+geom_point(aes(x=ElderlyRate,y=lethality))
```

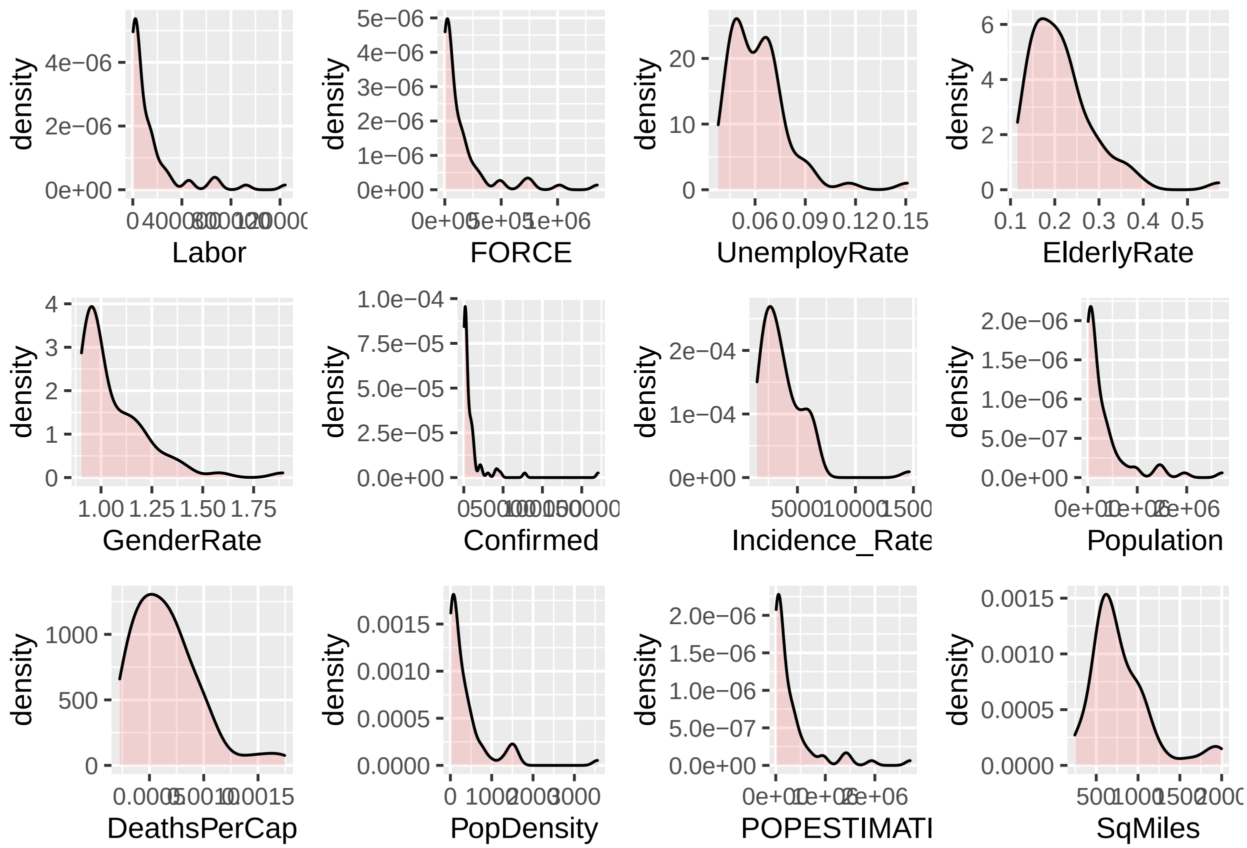



```r
#drop catagorical variables
df<-mydata[,-c(1,13,14,15,20,22)]
df<-df[,-c(10,11,12,17,16)]
#display descriptive statistics of all variables
summary(df)
```

```
##       FORCE              Labor          UnemploymentPOP      UnemployRate
##  Min.   :   2480   Min.   :   2375   Min.   :   103.0   Min.   :0.03800
##  1st Qu.:  11196   1st Qu.:  10672   1st Qu.:   524.5   1st Qu.:0.04900
##  Median :  46791   Median :  43385   Median :  3307.0   Median :0.06100
##  Mean   : 151299   Mean   : 139667   Mean   : 11631.5   Mean   :0.06206
##  3rd Qu.: 163012   3rd Qu.: 150547   3rd Qu.: 11126.5   3rd Qu.:0.06950
##  Max.   :1351810   Max.   :1242545   Max.   :109265.0   Max.   :0.15100
##   POPESTIMATE      AGE65PLUS_TOT     POPEST_MALE        POPEST_FEM
##  Min.   :   8236   Min.   :  1187   Min.   :   4931   Min.   :   3191
##  1st Qu.:  27704   1st Qu.:  4586   1st Qu.:  14809   1st Qu.:  13084
##  Median : 124995   Median : 32012   Median :  62419   Median :  62576
##  Mean   : 312890   Mean   : 62802   Mean   : 152983   Mean   : 159907
##  3rd Qu.: 353481   3rd Qu.: 82993   3rd Qu.: 171963   3rd Qu.: 181518
##  Max.   :2713295   Max.   :431554   Max.   :1318403   Max.   :1394892
##   ElderlyRate         Latitude       Longitude         Confirmed
##  Min.   :0.1158   Min.   :25.21   Min.   :-87.37   Min.   :   513
##  1st Qu.:0.1598   1st Qu.:27.59   1st Qu.:-83.54   1st Qu.:  1286
##  Median :0.2000   Median :29.46   Median :-82.32   Median :  2840
##  Mean   :0.2163   Mean   :28.94   Mean   :-82.70   Mean   : 10558
##  3rd Qu.:0.2418   3rd Qu.:30.25   3rd Qu.:-81.34   3rd Qu.:  9262
```

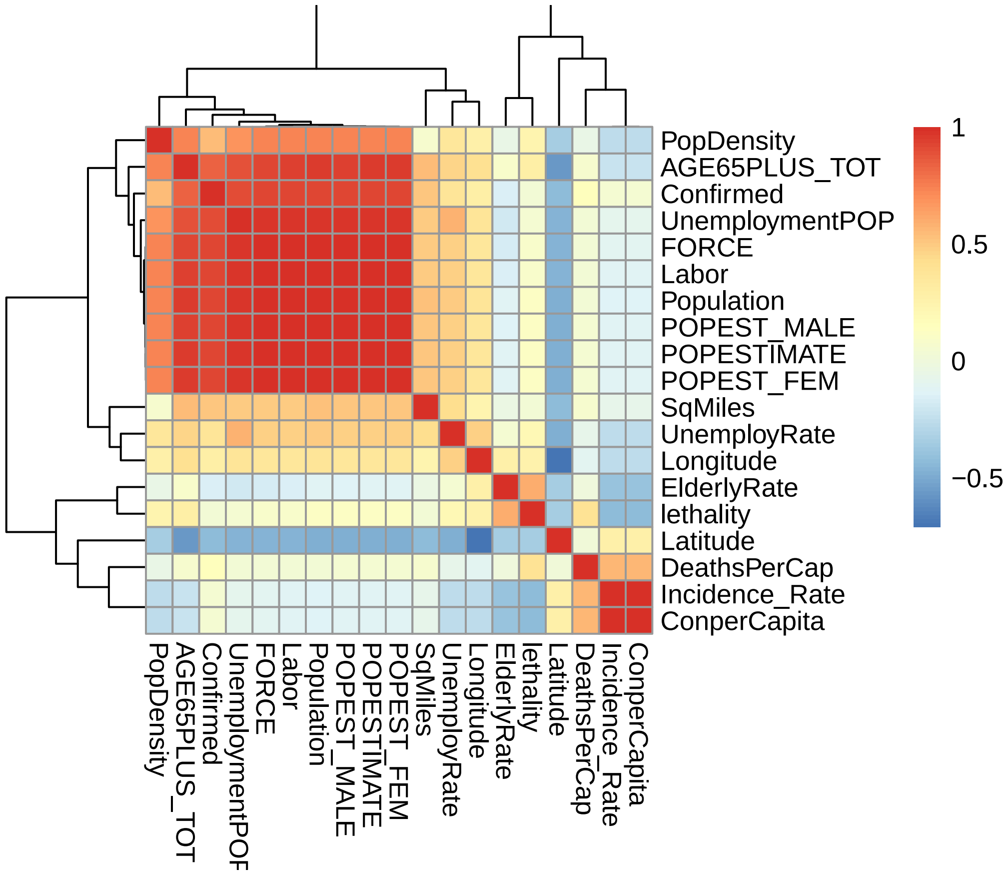
```
##  Max.   :0.5707   Max.   :30.87   Max.   :-80.43   Max.   :170882
##  Incidence_Rate    lethality         Population       ConperCapita
##  Min.   : 1521   Min.   :0.006369   Min.   :   8354   Min.   :0.01521
##  1st Qu.: 2413   1st Qu.:0.012473   1st Qu.:  28706   1st Qu.:0.02413
##  Median : 3120   Median :0.018158   Median : 132420   Median :0.03120
##  Mean   : 3694   Mean   :0.019807   Mean   : 320563   Mean   :0.03694
##  3rd Qu.: 4676   3rd Qu.:0.025423   3rd Qu.: 371435   3rd Qu.:0.04676
##  Max.   :14640   Max.   :0.043913   Max.   :2716940   Max.   :0.14640
##   DeathsPerCap          SqMiles          PopDensity
##  Min.   :0.0002259   Min.   : 243.6   Min.   :   9.998
##  1st Qu.:0.0004214   1st Qu.: 565.7   1st Qu.:  50.243
##  Median :0.0006027   Median : 696.0   Median : 192.613
##  Mean   :0.0006524   Mean   : 800.4   Mean   : 383.859
##  3rd Qu.:0.0008150   3rd Qu.: 960.8   3rd Qu.: 440.447
##  Max.   :0.0017452   Max.   :1998.3   Max.   :3560.979
```

```r
library(gridExtra)
p1<-ggplot(mydata,aes(x=Labor))+geom_density(colour="black", fill="#FF6666",alpha=0.2)
p2<-ggplot(mydata,aes(x=FORCE))+geom_density(color="black",alpha=.2, fill="#FF6666")
p3<-ggplot(mydata,aes(x=UnemployRate))+geom_density(color="black",alpha=.2, fill="#FF6666")
p4<-ggplot(mydata,aes(x=ElderlyRate))+geom_density(color="black",alpha=.2, fill="#FF6666")
p5<-ggplot(mydata,aes(x=GenderRate))+geom_density(color="black",alpha=.2, fill="#FF6666")
p6<-ggplot(mydata,aes(x=Confirmed))+geom_density(color="black",alpha=.2, fill="#FF6666")
p7<-ggplot(mydata,aes(x=Incidence_Rate))+geom_density(color="black",alpha=.2, fill="#FF6666")
p8<-ggplot(mydata,aes(x=Population))+geom_density(color="black",alpha=.2, fill="#FF6666")
p9<-ggplot(mydata,aes(x=DeathsPerCap))+geom_density(color="black",alpha=.2, fill="#FF6666")
p10<-ggplot(mydata,aes(x=PopDensity))+geom_density(color="black",alpha=.2, fill="#FF6666")
p11<-ggplot(mydata,aes(x=POPESTIMATE))+geom_density(color="black",alpha=.2, fill="#FF6666")
p12<-ggplot(mydata,aes(x=SqMiles))+geom_density(color="black",alpha=.2, fill="#FF6666")
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,ncol=4)
```



display the correlation within variables

```
library(pheatmap)
#create matrix of correlation
matrix<-cor(df)
#display the heatmap
pheatmap(matrix,cellwidth=10,cellheight =10)
```




```r
#Playing around with Model
#COVID has a high death rate among older individuals, testing it
Deaths = lm((Deaths/Confirmed) ~ ElderlyRate, mydata)
summary(Deaths)
```

```
##
## Call:
## lm(formula = (Deaths/Confirmed) ~ ElderlyRate, data = mydata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.016445 -0.005707  0.001046  0.004905  0.017184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.004038   0.002766    1.46    0.149
## ElderlyRate 0.072906   0.012051    6.05 7.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007586 on 65 degrees of freedom
## Multiple R-squared:  0.3602, Adjusted R-squared:  0.3504
## F-statistic:  36.6 on 1 and 65 DF,  p-value: 7.945e-08
```

```r
#Testing Population Density as an added factor, since it would be more likely to be transmitted
Deaths = lm((Deaths/Confirmed) ~ ElderlyRate + PopDensity , mydata)
summary(Deaths)
```

```
##
## Call:
## lm(formula = (Deaths/Confirmed) ~ ElderlyRate + PopDensity, data = mydata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0164348 -0.0052205 -0.0004736  0.0043705  0.0169637
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.968e-03  2.721e-03   0.723  0.47216
## ElderlyRate 7.464e-02  1.145e-02   6.521 1.28e-08 ***
## PopDensity  4.413e-06  1.538e-06   2.870  0.00556 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007197 on 64 degrees of freedom
## Multiple R-squared:  0.4332, Adjusted R-squared:  0.4155
## F-statistic: 24.46 on 2 and 64 DF,  p-value: 1.289e-08
```

```r
# P Values are well below 0.05, meaning these two factors are significant in predicting Death Rates

#Adding a third factor, UnemployRate
Deaths = lm((Deaths/Confirmed) ~ ElderlyRate + PopDensity + UnemployRate, mydata)
summary(Deaths)
```

```
##
## Call:
## lm(formula = (Deaths/Confirmed) ~ ElderlyRate + PopDensity +
##     UnemployRate, data = mydata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0164237 -0.0049047 -0.0002782  0.0049368  0.0164128
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.836e-04  3.892e-03  -0.150   0.8813
## ElderlyRate   7.386e-02  1.149e-02   6.427 1.97e-08 ***
## PopDensity    3.851e-06  1.657e-06   2.324   0.0233 *
## UnemployRate  4.732e-02  5.154e-02   0.918   0.3620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007205 on 63 degrees of freedom
## Multiple R-squared:  0.4407, Adjusted R-squared:  0.414
## F-statistic: 16.54 on 3 and 63 DF,  p-value: 4.883e-08
```

```r
#After testing it, UnemployRate has a P Value of .36, next model it will be removed

#Showing "Leakage"
Deaths = lm((Deaths/Confirmed) ~ ElderlyRate + PopDensity + lethality, mydata)
summary(Deaths)
```

```
##
## Call:
## lm(formula = (Deaths/Confirmed) ~ ElderlyRate + PopDensity +
##     lethality, data = mydata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -5.446e-10 -2.722e-10  4.169e-11  2.512e-10  5.092e-10
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.055e-10  1.178e-10  8.960e-01    0.374
## ElderlyRate -8.967e-10  6.365e-10 -1.409e+00    0.164
## PopDensity  -2.986e-14  7.041e-14 -4.240e-01    0.673
## lethality    1.000e+00  5.388e-09  1.856e+08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.102e-10 on 63 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.026e+16 on 3 and 63 DF,  p-value: < 2.2e-16
```

```r
#As seen here, adding lethality results in a perfect R square, since it goes hand in hand with Death Ra

#Many variables
Deaths = lm((Deaths/Confirmed) ~ UnemployRate + ElderlyRate + PopDensity, mydata)
summary(Deaths)
```

```
##
## Call:
## lm(formula = (Deaths/Confirmed) ~ UnemployRate + ElderlyRate +
##     PopDensity, data = mydata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0164237 -0.0049047 -0.0002782  0.0049368  0.0164128
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.836e-04  3.892e-03  -0.150   0.8813
## UnemployRate  4.732e-02  5.154e-02   0.918   0.3620
## ElderlyRate   7.386e-02  1.149e-02   6.427 1.97e-08 ***
## PopDensity    3.851e-06  1.657e-06   2.324   0.0233 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007205 on 63 degrees of freedom
```

```
## Multiple R-squared:  0.4407, Adjusted R-squared:  0.414
## F-statistic: 16.54 on 3 and 63 DF,  p-value: 4.883e-08
```

```r
#With our current set of data, we found that the two factors below are the best predictors
Deaths = lm((Deaths/Confirmed) ~ ElderlyRate + PopDensity, mydata)
summary(Deaths)
```

```
##
## Call:
## lm(formula = (Deaths/Confirmed) ~ ElderlyRate + PopDensity, data = mydata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0164348 -0.0052205 -0.0004736  0.0043705  0.0169637
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.968e-03  2.721e-03   0.723  0.47216
## ElderlyRate 7.464e-02  1.145e-02   6.521 1.28e-08 ***
## PopDensity  4.413e-06  1.538e-06   2.870  0.00556 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007197 on 64 degrees of freedom
## Multiple R-squared:  0.4332, Adjusted R-squared:  0.4155
## F-statistic: 24.46 on 2 and 64 DF,  p-value: 1.289e-08
```

```r
lithalitymod<- lm(log(mydata$lethality)~log(mydata$ElderlyRate)+log(mydata$PopDensity))
summary(lithalitymod)
```

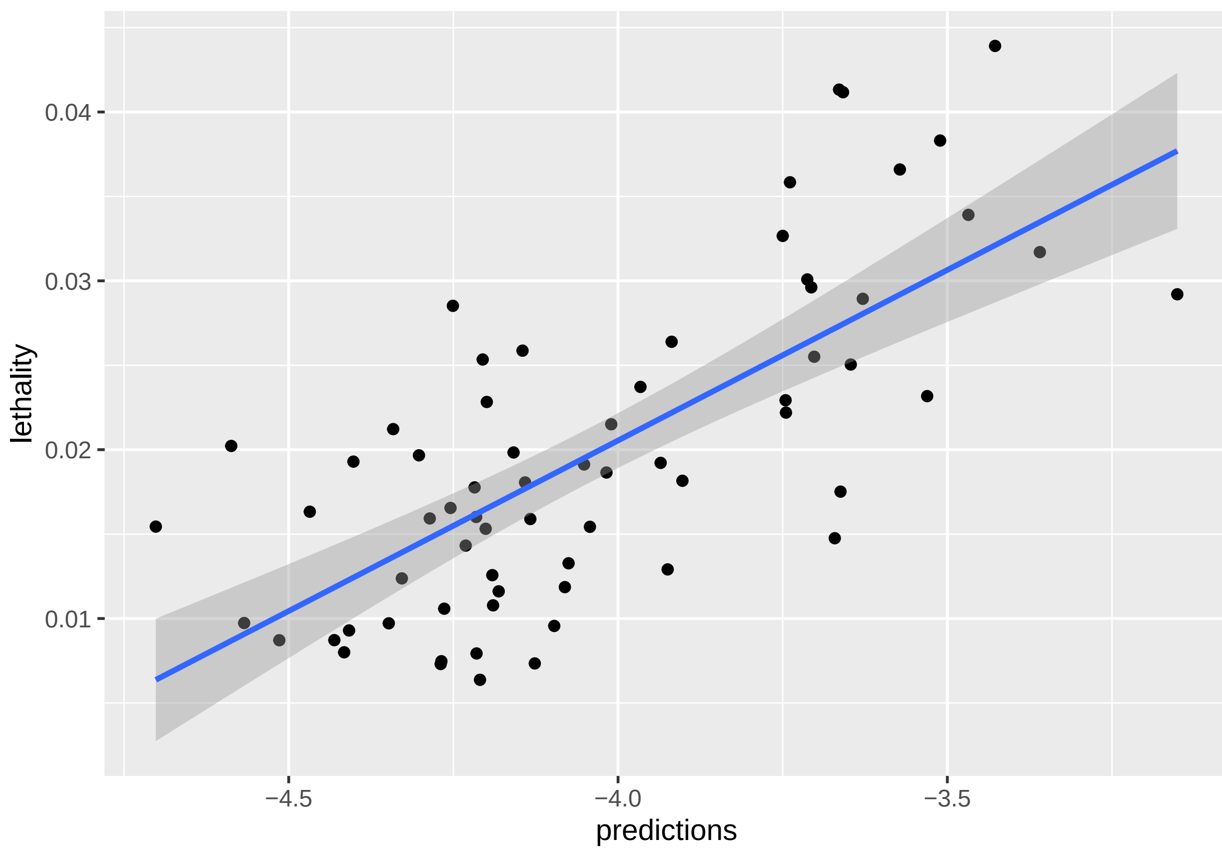
```
##
## Call:
## lm(formula = log(mydata$lethality) ~ log(mydata$ElderlyRate) +
##     log(mydata$PopDensity))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84666 -0.27254  0.02212  0.25662  0.69361
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.4583     0.2939 -11.765  < 2e-16 ***
## log(mydata$ElderlyRate)   0.8064     0.1444   5.586 5.10e-07 ***
## log(mydata$PopDensity)    0.1384     0.0321   4.312 5.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3688 on 64 degrees of freedom
## Multiple R-squared:  0.4573, Adjusted R-squared:  0.4403
## F-statistic: 26.97 on 2 and 64 DF,  p-value: 3.204e-09
```

```r
y_hat <- predict(lithalitymod, se.fit = TRUE)
mydata$predictions <- y_hat$fit
mydata$residuals <- lithalitymod$residuals
ObsNum <- 1:(length(mydata$lethality))
mydata$ObsNum <- ObsNum



#Assumption 1: There is a linear relationship
#between the dependent variable Y and the independent variable X
ggplot(data = mydata, aes(predictions, lethality)) +
geom_point()+geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

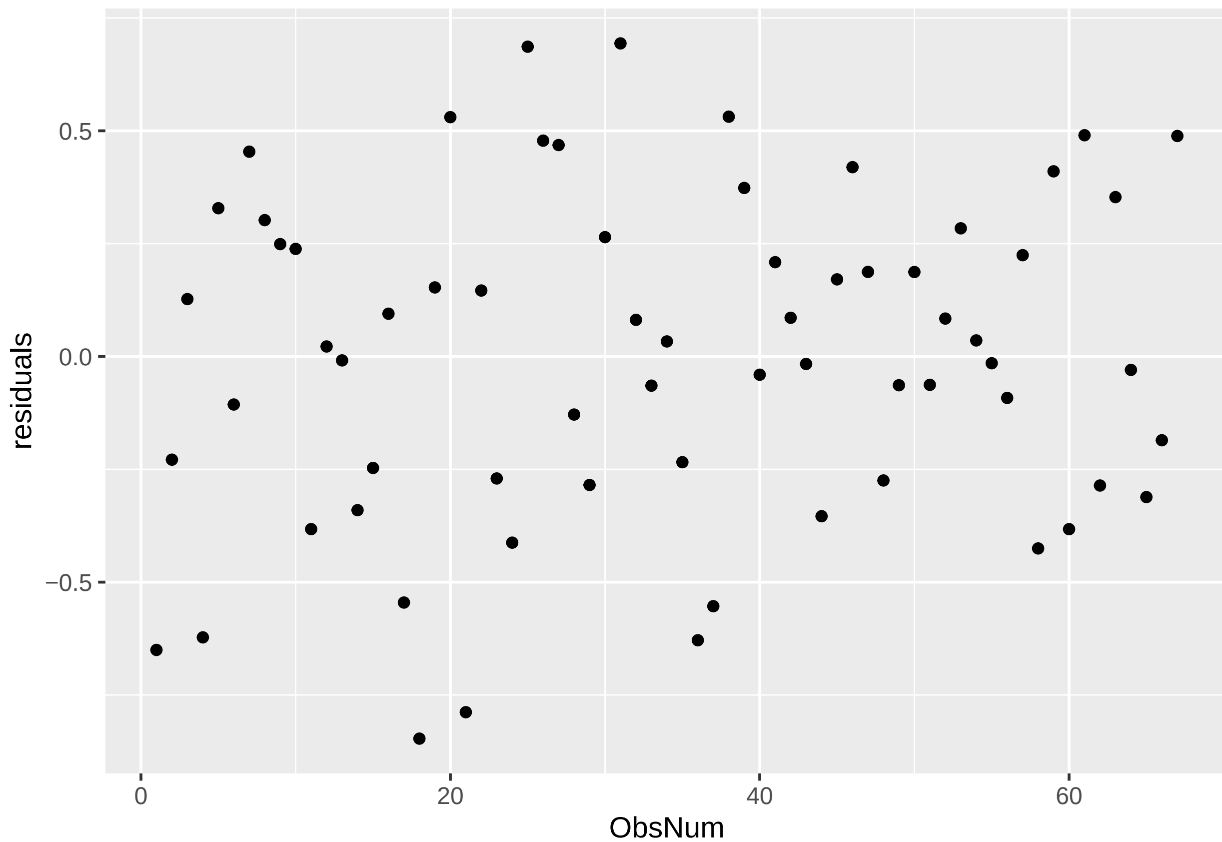



```r
#Assumption 2: Residuals are independent.
residualmod <- lm(residuals~ObsNum, mydata)
summary(residualmod)
```

```
##
## Call:
## lm(formula = residuals ~ ObsNum, data = mydata)
##
## Residuals:
```

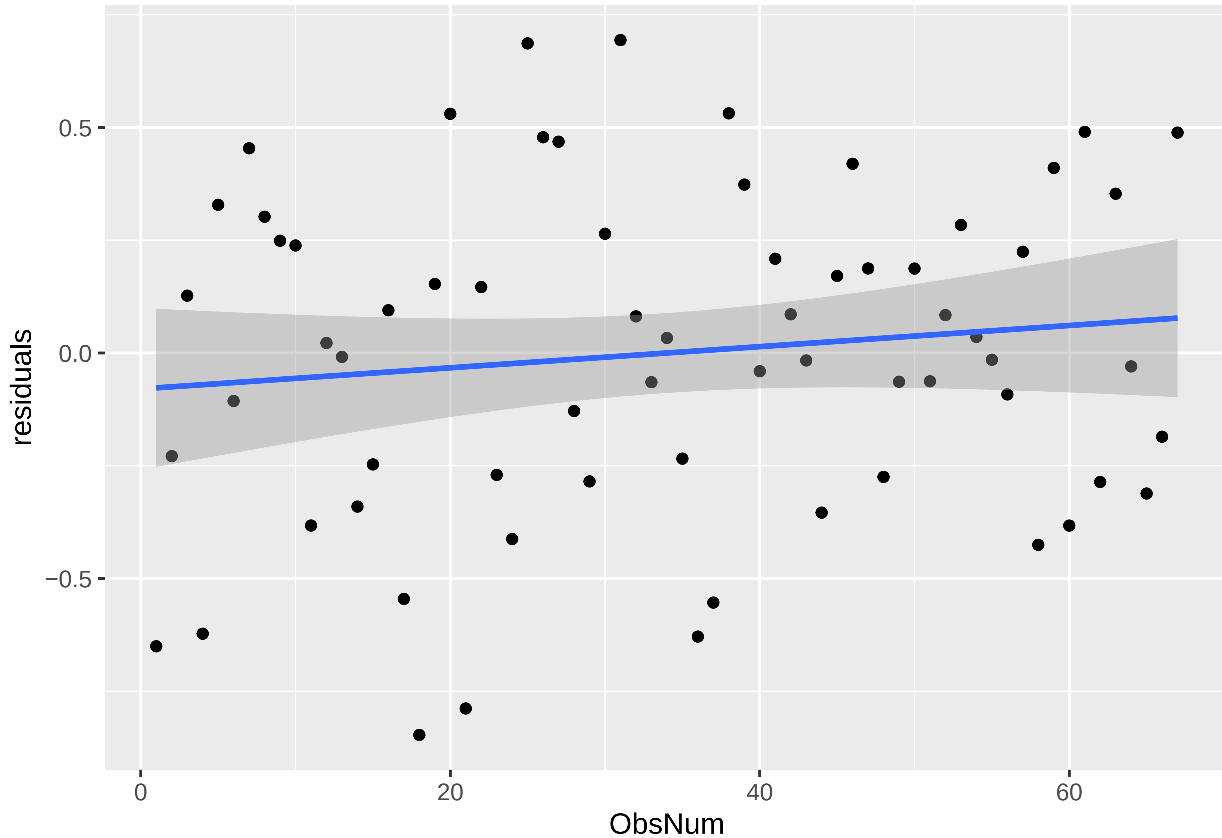
```
##      Min       1Q   Median       3Q      Max
## -0.80916 -0.26681  0.03328  0.27937  0.70738
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.079696   0.089714  -0.888    0.378
## ObsNum       0.002344   0.002294   1.022    0.311
##
## Residual standard error: 0.3631 on 65 degrees of freedom
## Multiple R-squared:  0.01581,    Adjusted R-squared:  0.000673
## F-statistic: 1.044 on 1 and 65 DF,  p-value: 0.3106
```

```
ggplot(data = mydata, aes(ObsNum, residuals))+geom_point()
```




```
ggplot(data = mydata, aes(ObsNum, residuals))+geom_point()+geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



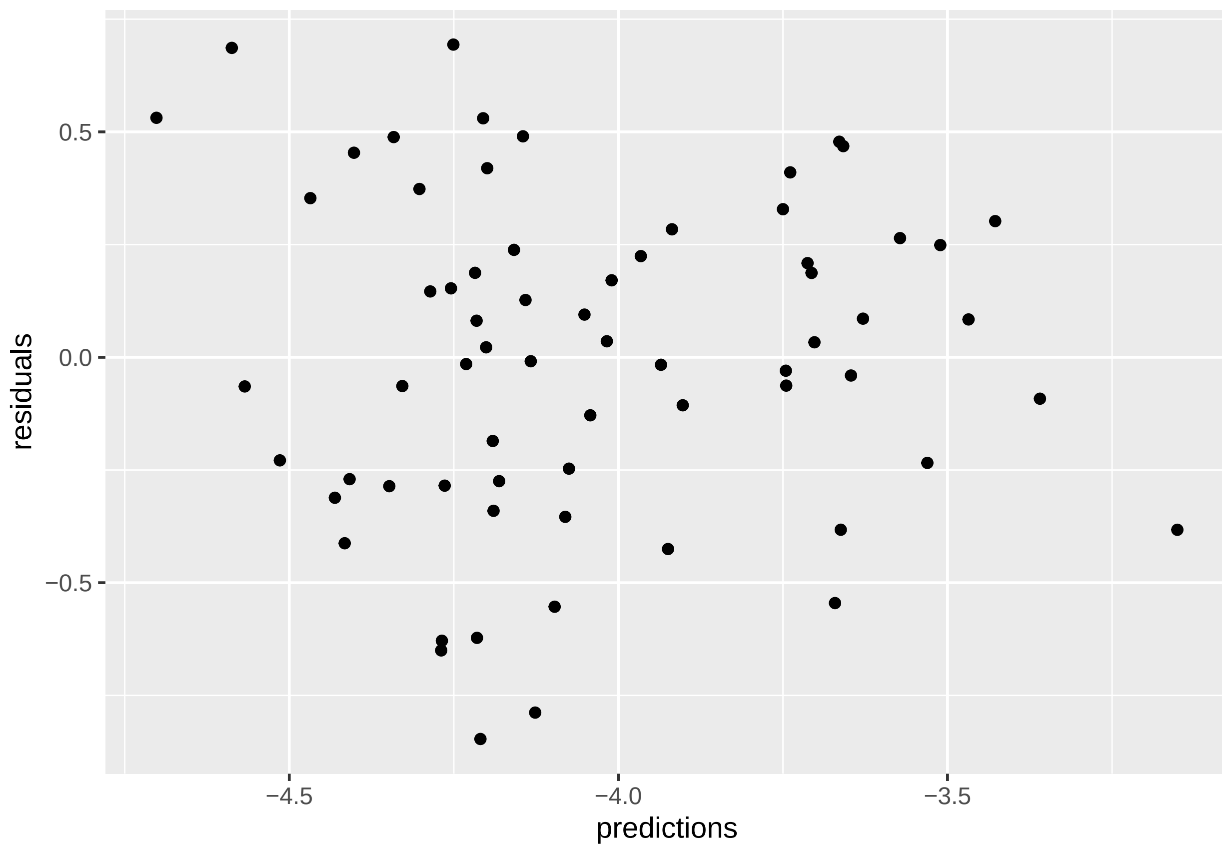

```
#Assumption 3: Constant variance
library(car)
ncvTest(lithalitymod)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.940014, Df = 1, p = 0.08641
```

```
bptest(lithalitymod, varformula = NULL, studentize = TRUE, data = lithalitymod$model())
```

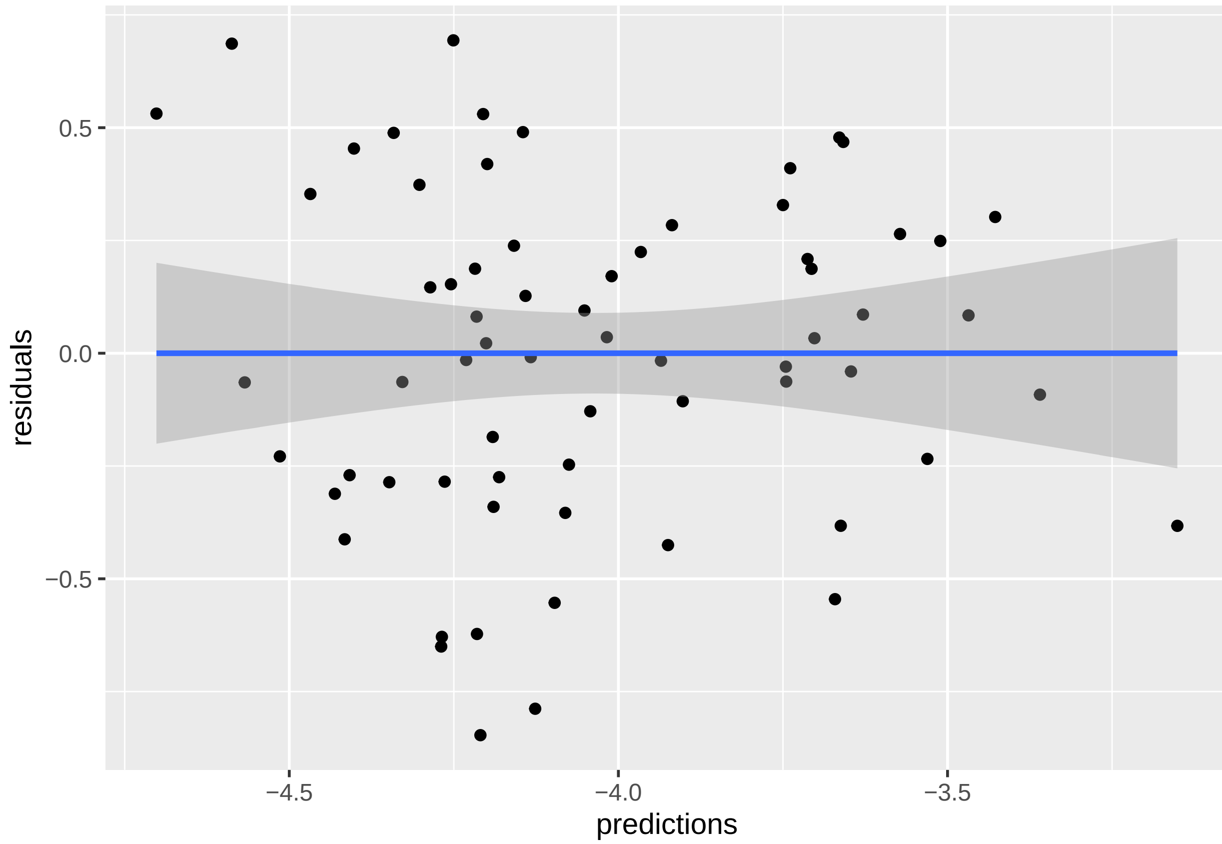
```
##
##  studentized Breusch-Pagan test
##
## data:  lithalitymod
## BP = 9.4722, df = 2, p-value = 0.008773
```

```r
ggplot(mydata, aes(predictions, residuals))+geom_point()
```

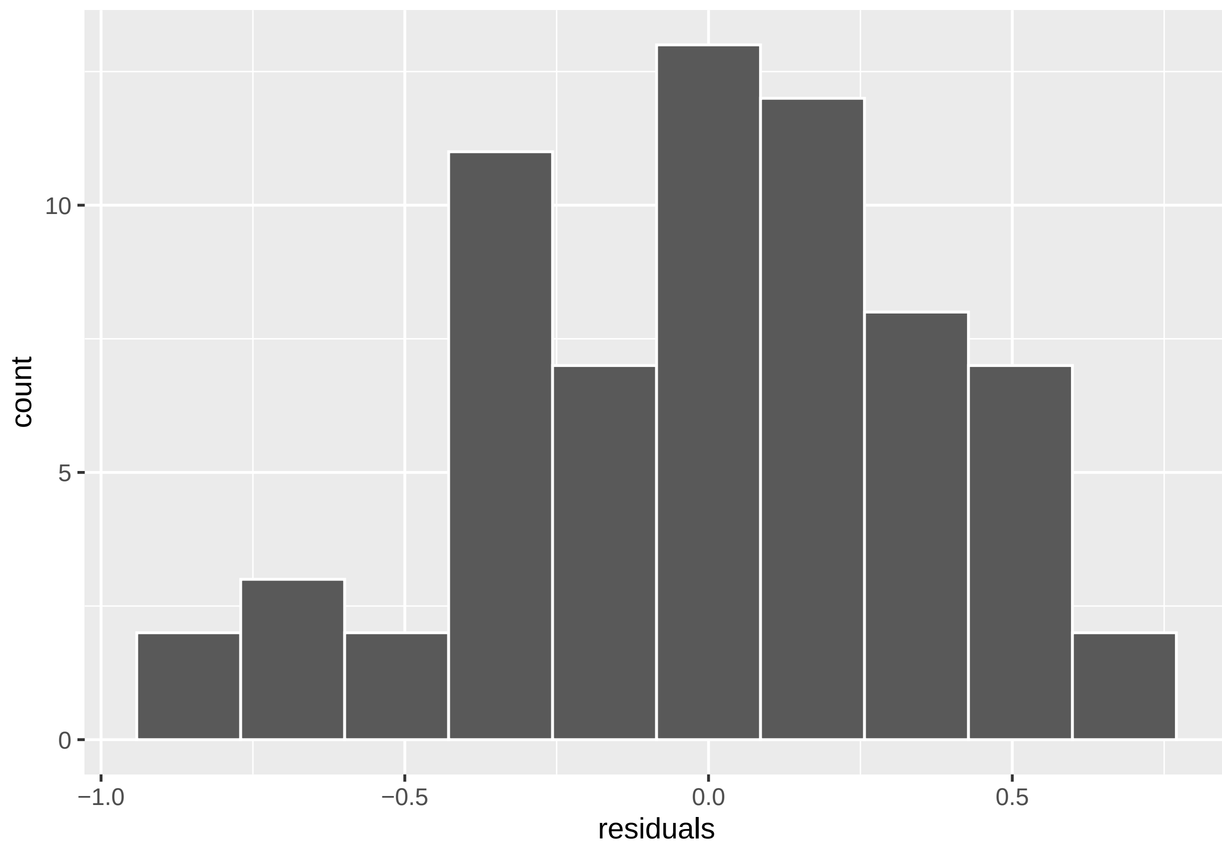


```r
ggplot(mydata, aes(predictions, residuals))+geom_point()+geom_smooth(method = "lm")
```

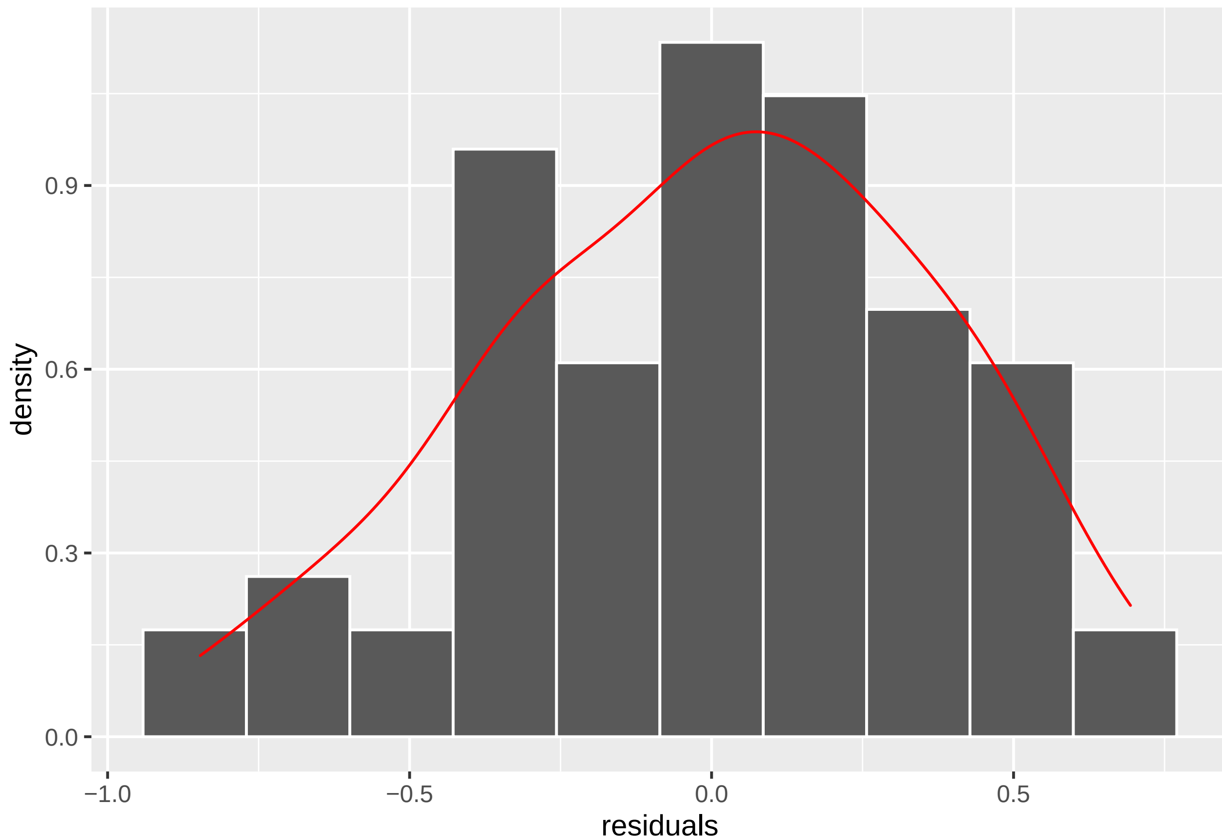
```
## `geom_smooth()` using formula 'y ~ x'
```



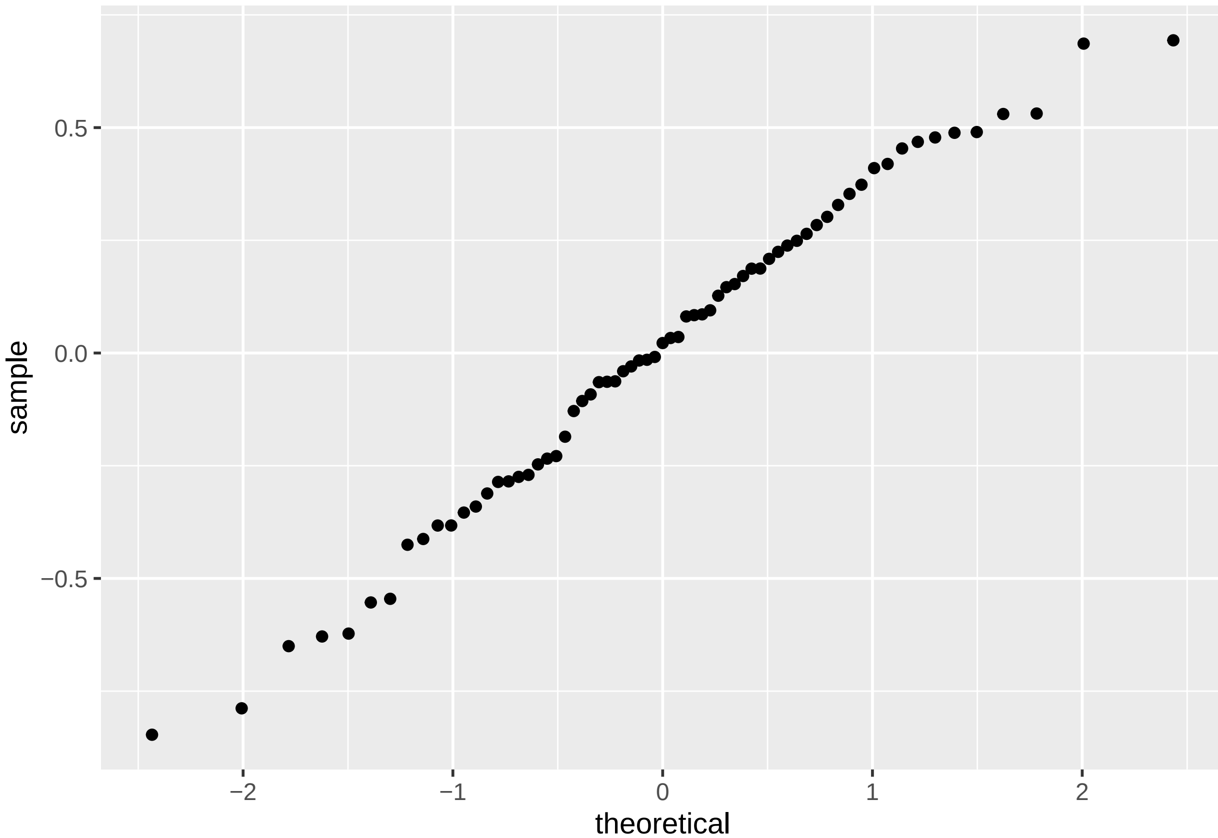

```
#Normality of residuals
ggplot(data = mydata, aes(x=residuals)) + geom_histogram(bins=10, col="white")
```



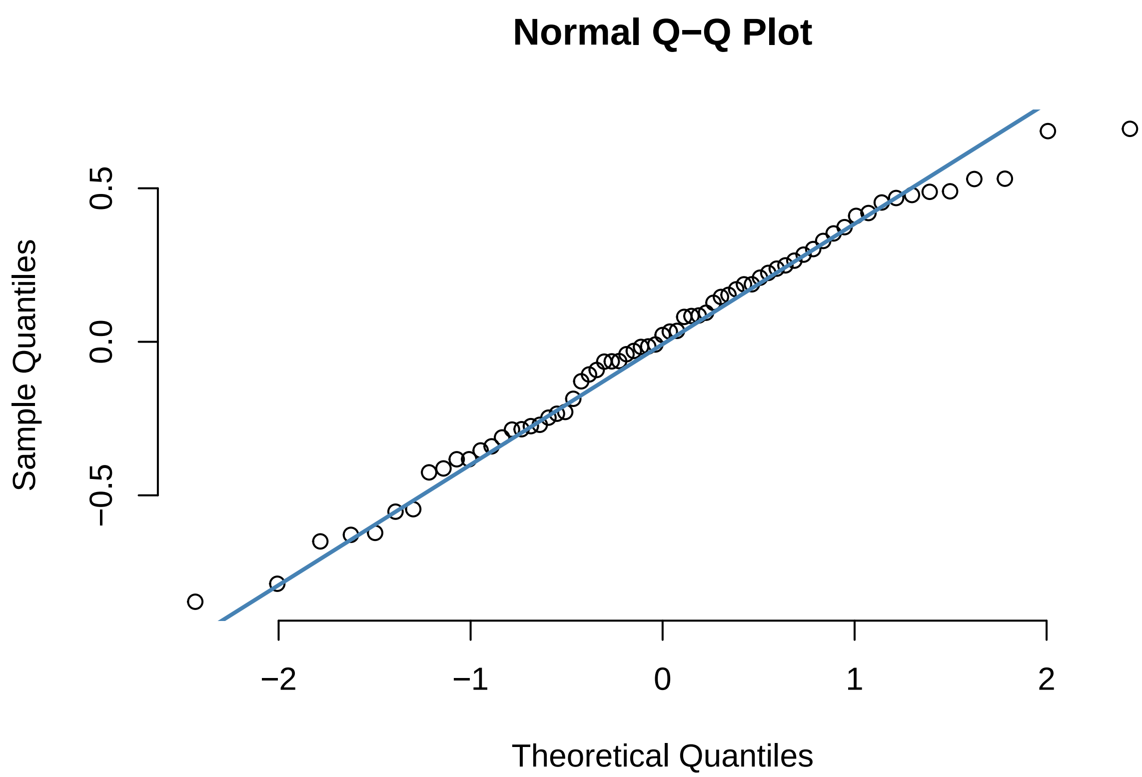

```
ggplot(data = mydata, aes(x=residuals)) +
  geom_histogram(aes(y=..density..), bins=10, col="white") +
  geom_density(aes(y=..density..), colour="red")
```




```
#Normality of residuals cont
ggplot(mydata, aes(sample=residuals))+
stat_qq()
```




```
# base r qqplot
layout(matrix(c(1,1)))
qqnorm(mydata$residuals, pch = 1, frame = FALSE)
qqline(mydata$residuals, col = "steelblue", lwd = 2)
```




```
# statistical test for normality
shapiro.test(mydata$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mydata$residuals
## W = 0.98354, p-value = 0.5194
```

```
ad.test(mydata$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  mydata$residuals
## A = 0.26226, p-value = 0.6935
```

```
#Check multicollinearity

vif(lithalitymod)
```

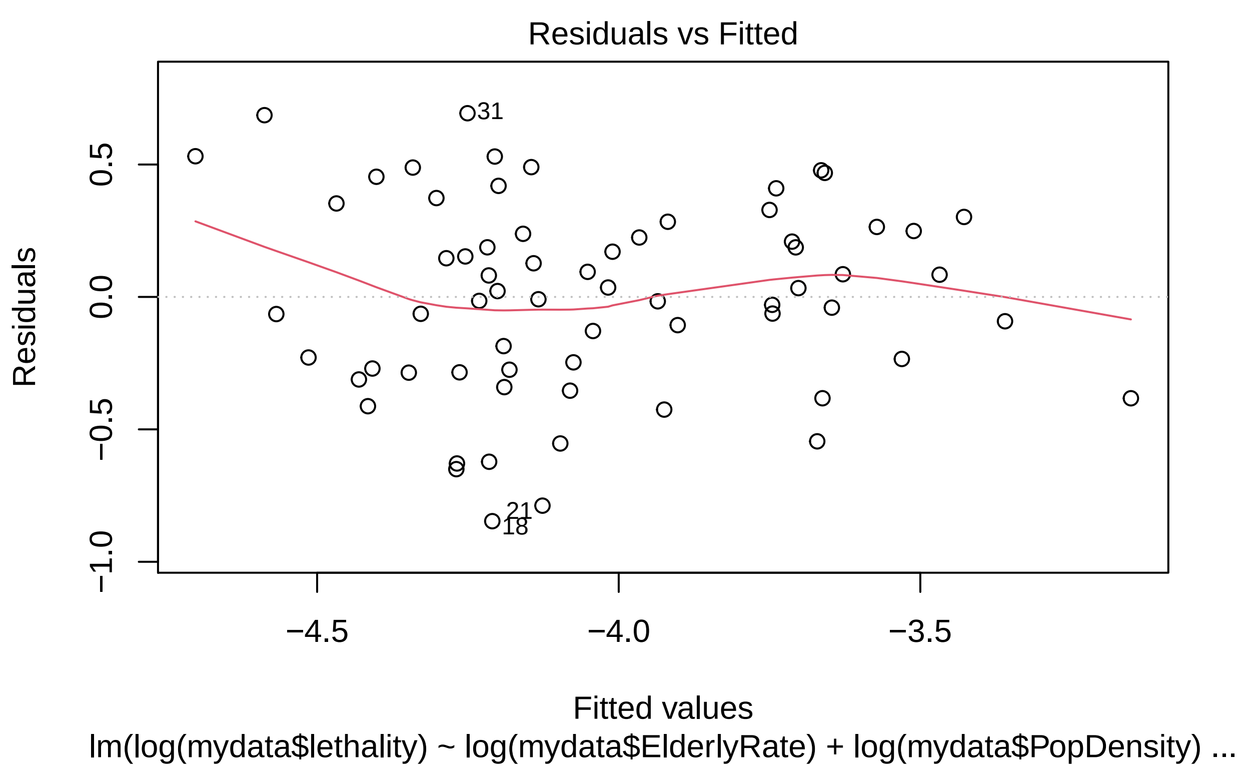
```
## log(mydata$ElderlyRate)  log(mydata$PopDensity)
##                1.006267                1.006267
```

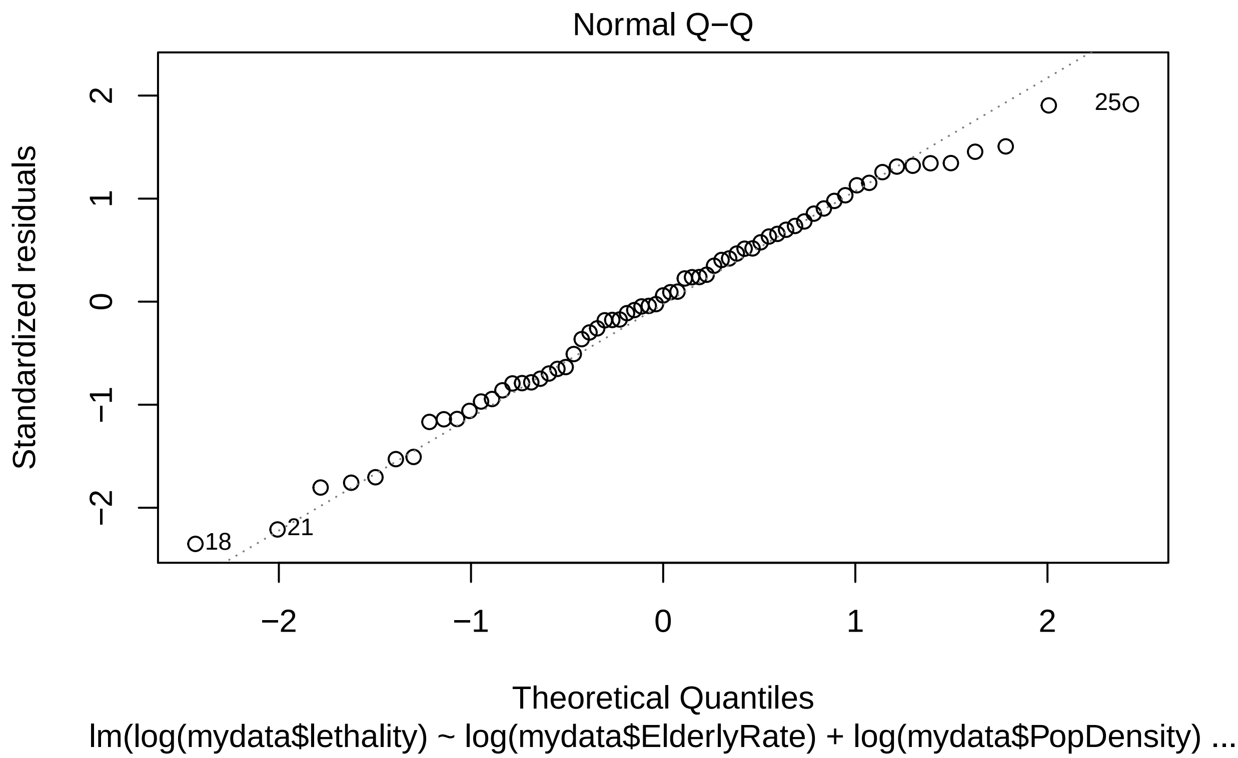
```r
#Check high leverage data points
cooksd <- data.frame(cooks.distance(lithalitymod))
cooks.distance(lithalitymod) > 4 / length(cooks.distance(lithalitymod))
```

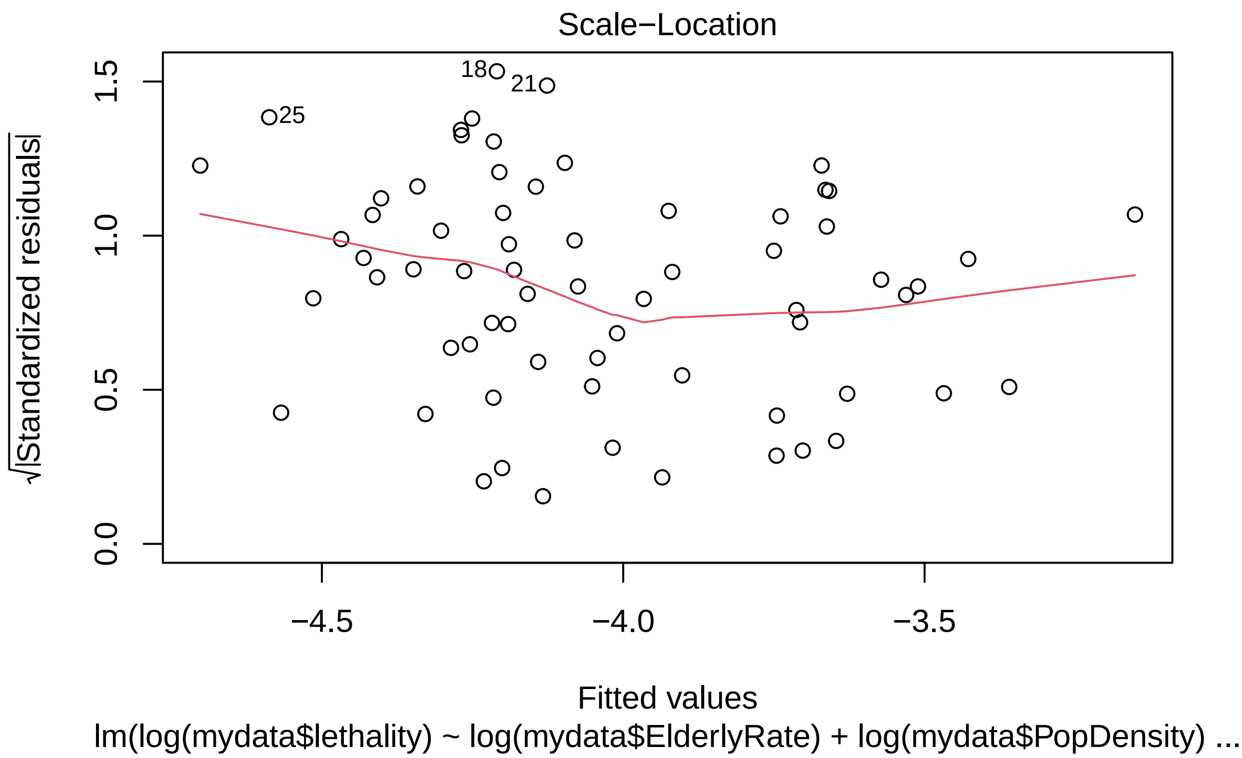
```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19    20    21    22    23    24    25    26
## FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
##    27    28    29    30    31    32    33    34    35    36    37    38    39
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
##    40    41    42    43    44    45    46    47    48    49    50    51    52
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    53    54    55    56    57    58    59    60    61    62    63    64    65
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##    66    67
## FALSE FALSE
```

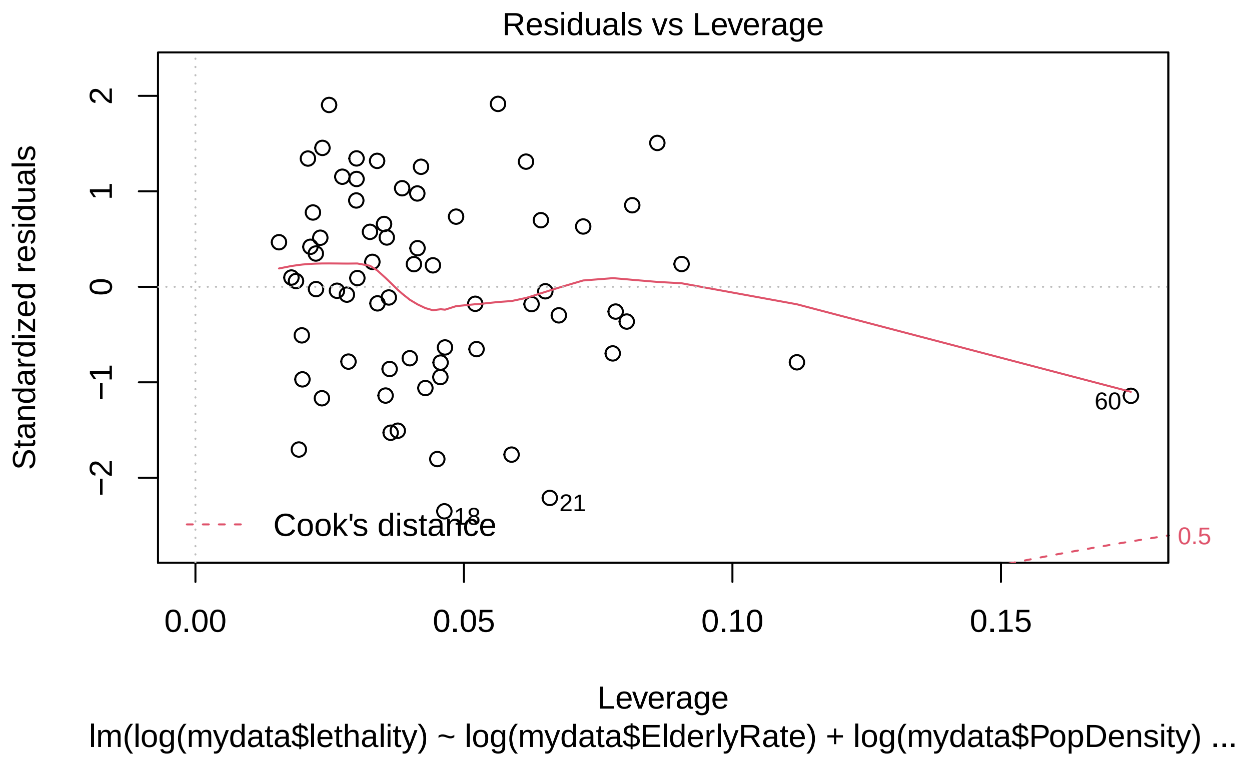
```r
cooksd$hiLev <- cooks.distance(lithalitymod) > 4 / length(cooks.distance(lithalitymod))
```

```r
#plot of model
plot(lithalitymod)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(mydata$lethality) ~ log(mydata$ElderlyRate) + log(mydata$PopDensity) ...
18
21
25
−2
−1
0
1
2


Scale−Location
18
21
25
1.5
1.0
0.5
0.0
√|Standardized residuals|
−4.5
−4.0
−3.5
Fitted values
lm(log(mydata$lethality) ~ log(mydata$ElderlyRate) + log(mydata$PopDensity) ...


Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
60
21
18
Leverage
lm(log(mydata$lethality) ~ log(mydata$ElderlyRate) + log(mydata$PopDensity) ...