Analysis

Dingyi Li

4/14/2021

Multiple Linear Regression

Data preview

Read in data

```
dt = read.csv("data&figures/dt.csv")
summary(dt)
```

```
##     State              County               HPI          Personal_Income
##  Length:2553        Length:2553        Min.   :  82.32   Min.   : 22440
##  Class :character   Class :character   1st Qu.: 183.00   1st Qu.: 38315
##  Mode  :character   Mode  :character   Median : 237.40   Median : 43554
##                                        Mean   : 308.90   Mean   : 45980
##                                        3rd Qu.: 362.41   3rd Qu.: 50399
##                                        Max.   :2266.07   Max.   :229825
##  Poverty_Percentage   Population      HighSchoolLess  HighSchoolOnly
##  Min.   : 2.70      Min.   :     728   Min.   : 1.40   Min.   : 7.80
##  1st Qu.:10.10      1st Qu.:   15232   1st Qu.: 8.30   1st Qu.:29.40
##  Median :13.00      Median :   31829   Median :11.40   Median :34.40
##  Mean   :13.77      Mean   :  119242   Mean   :12.51   Mean   :33.95
##  3rd Qu.:16.60      3rd Qu.:   80485   3rd Qu.:15.80   3rd Qu.:38.90
##  Max.   :38.20      Max.   :10039107   Max.   :46.70   Max.   :54.50
##   SomeCollege    BachelorAndHigher Unemployment_Rate
##  Min.   :11.20   Min.   : 7.20     Min.   : 1.600
##  1st Qu.:27.70   1st Qu.:15.80     1st Qu.: 3.100
##  Median :31.00   Median :20.10     Median : 3.700
##  Mean   :31.01   Mean   :22.53     Mean   : 3.924
##  3rd Qu.:34.20   3rd Qu.:26.90     3rd Qu.: 4.500
##  Max.   :47.30   Max.   :75.30     Max.   :18.300
```

Correlation Check

```
cor(scale(as.matrix(dt[,c(7,8,9,10)])))
```

Education parameters

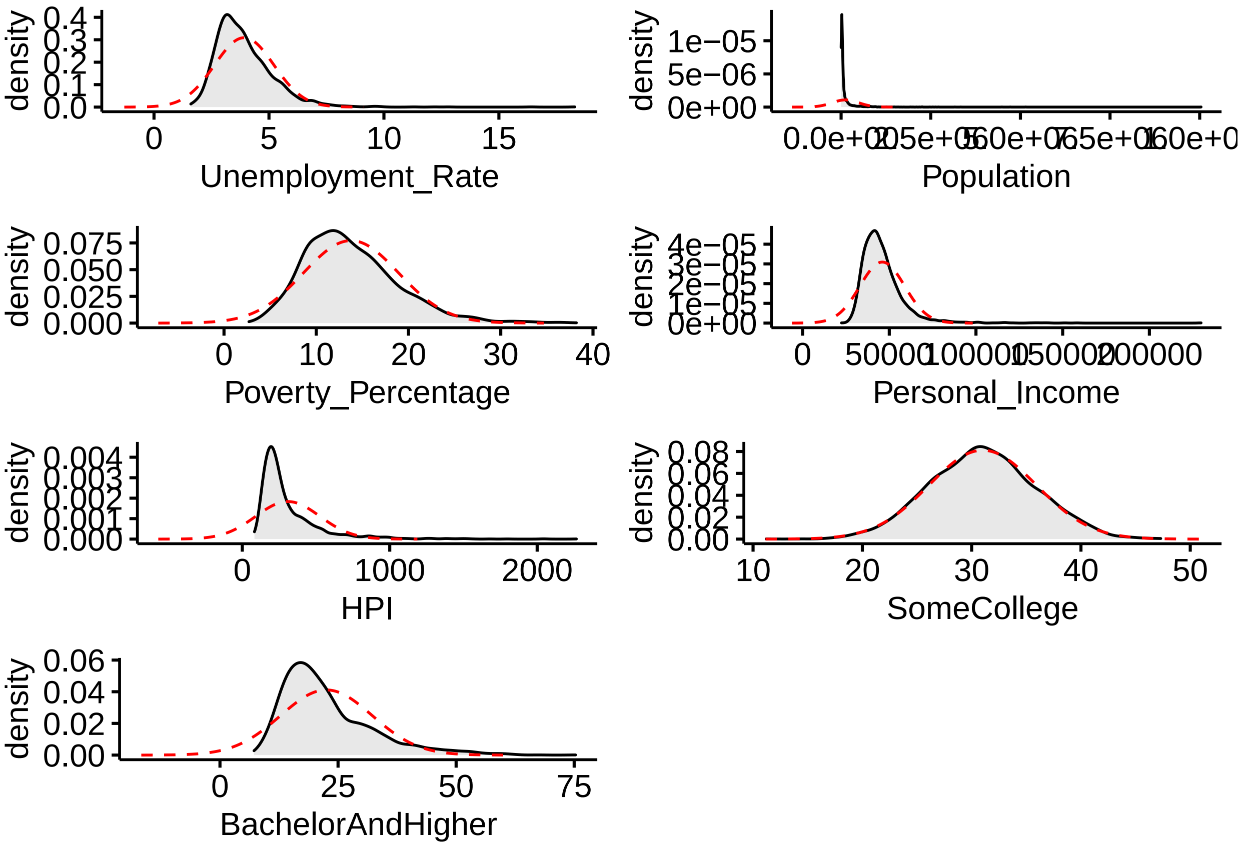
```
##                    HighSchoolLess HighSchoolOnly SomeCollege BachelorAndHigher
## HighSchoolLess          1.0000000      0.2825900 -0.39212427       -0.60271768
## HighSchoolOnly          0.2825900      1.0000000 -0.24778539       -0.79511586
## SomeCollege            -0.3921243     -0.2477854  1.00000000       -0.08951844
## BachelorAndHigher      -0.6027177     -0.7951159 -0.08951844        1.00000000
```

Histogram

```r
library(ggpubr)
```

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

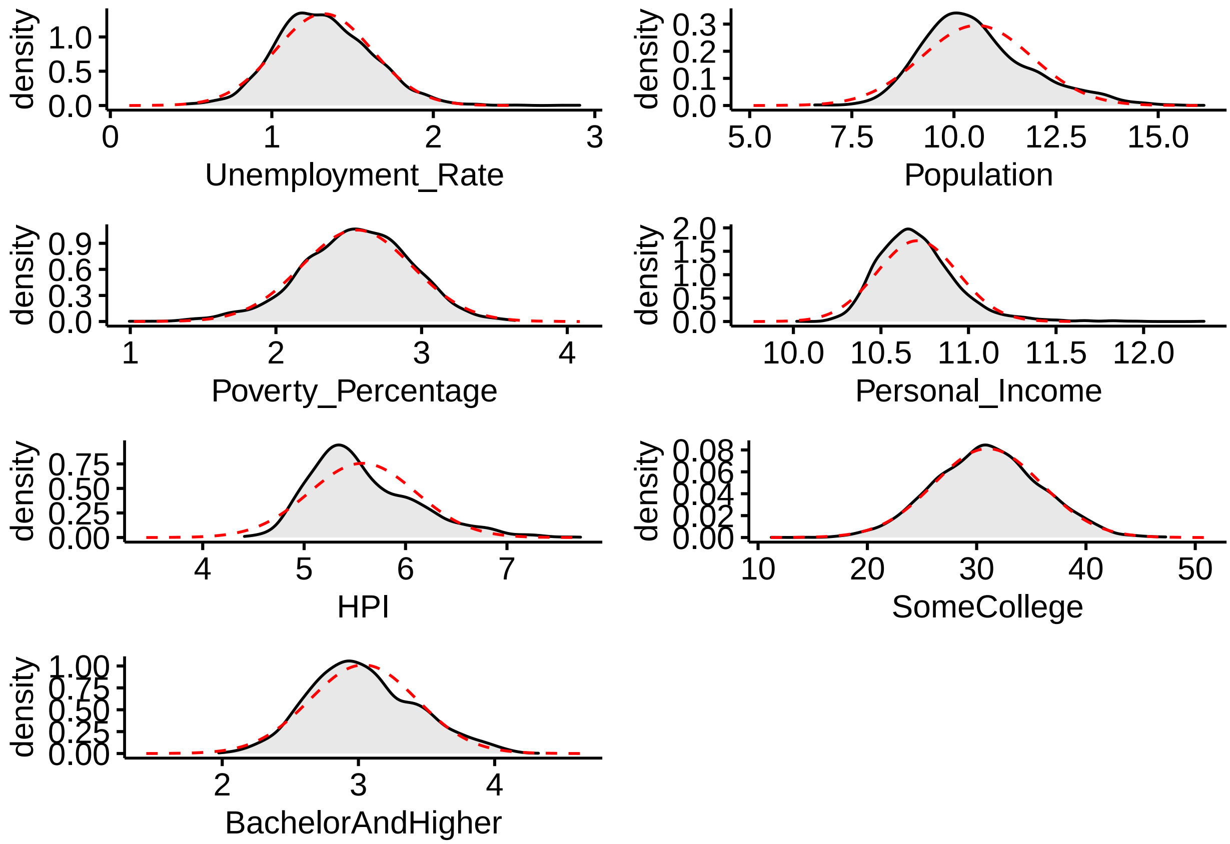
```r
a<-ggdensity(dt, x = "Unemployment_Rate", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
b<-ggdensity(dt, x = "Population", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
c<-ggdensity(dt, x = "Poverty_Percentage", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
d<-ggdensity(dt, x = "Personal_Income", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
e<-ggdensity(dt, x = "HPI", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
f<-ggdensity(dt, x = "SomeCollege", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
g<-ggdensity(dt, x = "BachelorAndHigher", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
ggarrange(a,b,c,d,e,f,g, ncol = 2, nrow = 4)
```



### Histogram for logtransformation

```
temp=dt
temp$HPI <- log(dt$HPI)
temp$Personal_Income <- log(dt$Personal_Income)
temp$Poverty_Percentage <- log(dt$Poverty_Percentage)
temp$Population <- log(dt$Population)
temp$HighSchoolLess <- log(dt$HighSchoolLess)
temp$BachelorAndHigher <- log(dt$BachelorAndHigher)
temp$Unemployment_Rate <- log(dt$Unemployment_Rate)
```

```
library(ggpubr)
a<-ggdensity(temp, x = "Unemployment_Rate", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
b<-ggdensity(temp, x = "Population", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
c<-ggdensity(temp, x = "Poverty_Percentage", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
d<-ggdensity(temp, x = "Personal_Income", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
e<-ggdensity(temp, x = "HPI", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
f<-ggdensity(temp, x = "SomeCollege", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
g<-ggdensity(temp, x = "BachelorAndHigher", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
ggarrange(a,b,c,d,e,f,g, ncol = 2, nrow = 4)
```



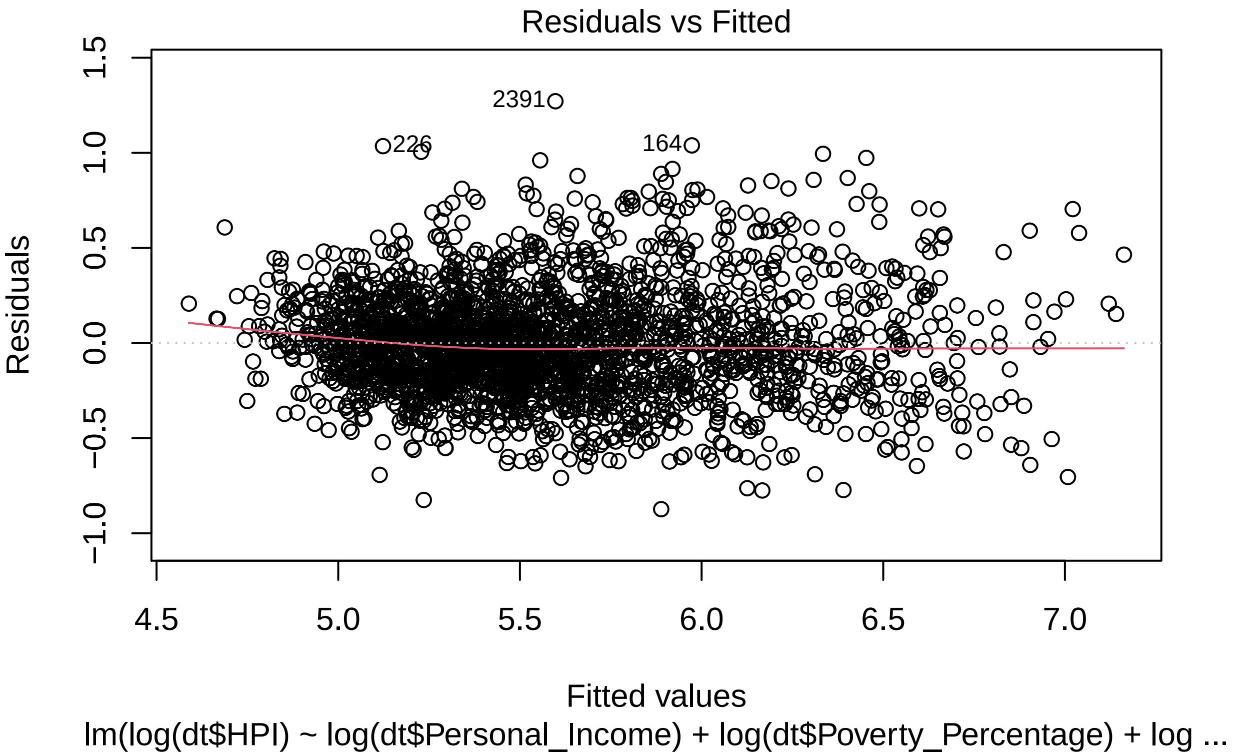
Model fitting

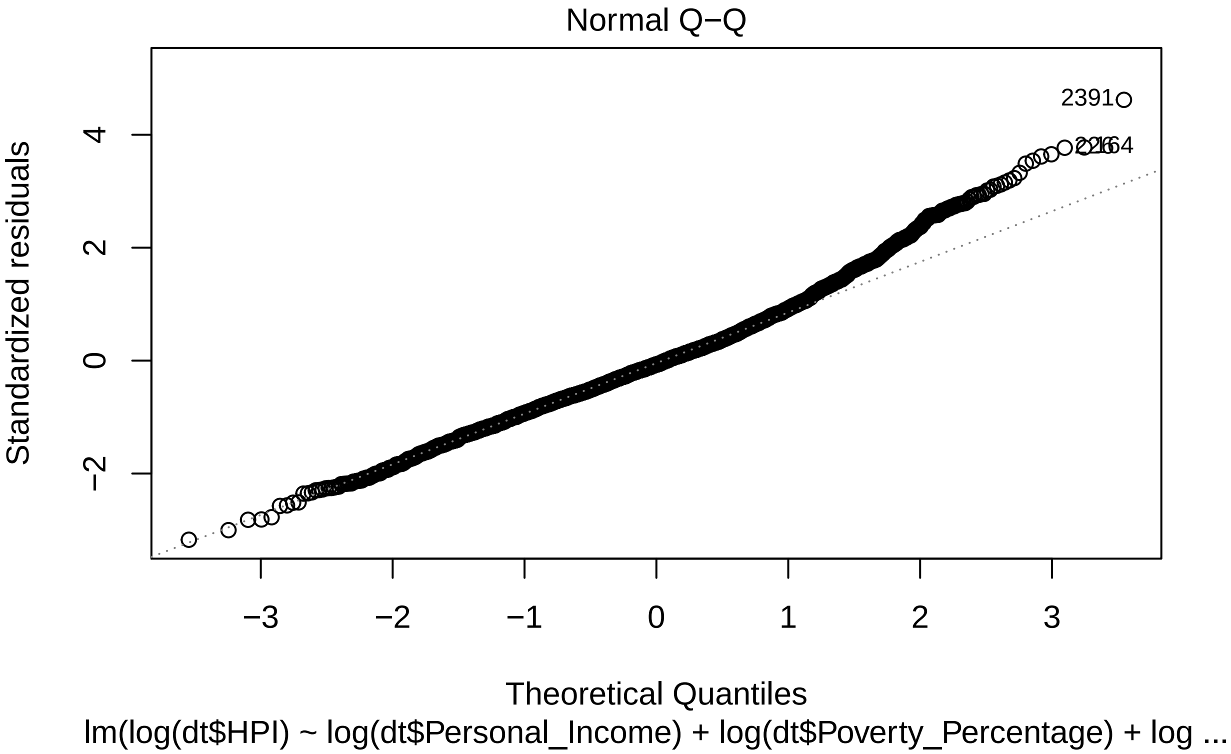
```
m1 = lm(log(dt$HPI)~log(dt$Personal_Income)+log(dt$Poverty_Percentage)+log(dt$Unemployment_Rate)+log(dt
summary(m1)
```

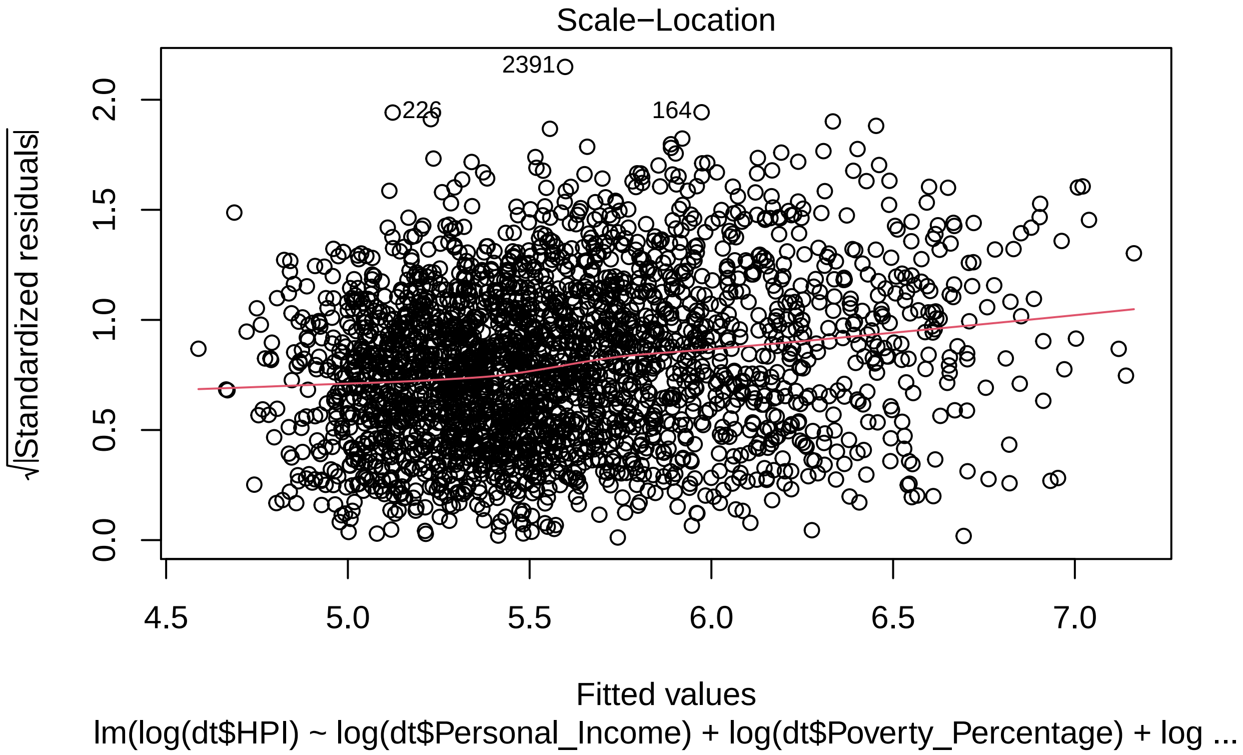
```
##
## Call:
## lm(formula = log(dt$HPI) ~ log(dt$Personal_Income) + log(dt$Poverty_Percentage) +
##     log(dt$Unemployment_Rate) + log(dt$Population) + dt$SomeCollege +
##     log(dt$BachelorAndHigher))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87316 -0.17859 -0.01902  0.15432  1.27125
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -3.860460   0.441084  -8.752  < 2e-16 ***
## log(dt$Personal_Income)     0.565650   0.039981  14.148  < 2e-16 ***
## log(dt$Poverty_Percentage) -0.071382   0.022242  -3.209  0.00135 **
## log(dt$Unemployment_Rate)   0.023098   0.022133   1.044  0.29678
## log(dt$Population)          0.237943   0.004936  48.203  < 2e-16 ***
## dt$SomeCollege              0.012922   0.001180  10.954  < 2e-16 ***
## log(dt$BachelorAndHigher)   0.204123   0.022826   8.942  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2755 on 2546 degrees of freedom
## Multiple R-squared:  0.7192, Adjusted R-squared:  0.7185
## F-statistic:  1087 on 6 and 2546 DF,  p-value: < 2.2e-16
```

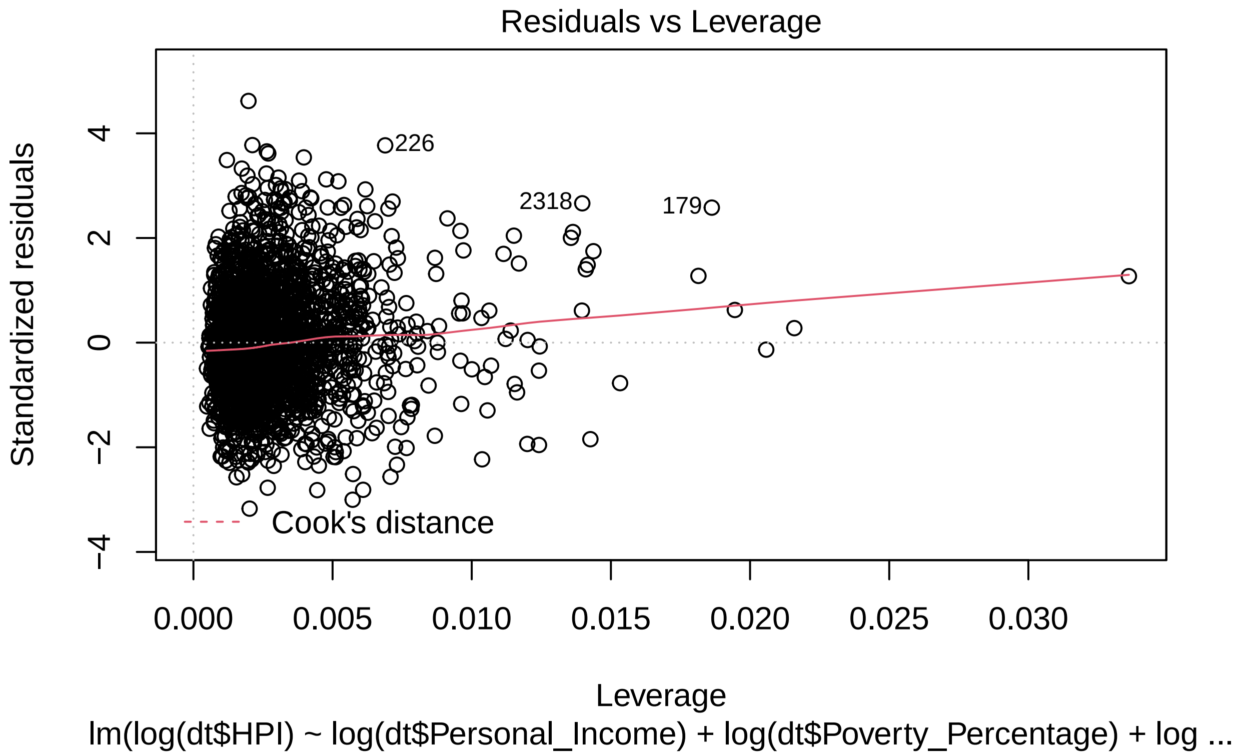
Diagnostic Plots

```r
plot(m1)
```




Normal Q−Q
2391
216
2164
Standardized residuals
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(dt$HPI) ~ log(dt$Personal_Income) + log(dt$Poverty_Percentage) + log ...


Scale−Location
2391
226
164
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
Fitted values
lm(log(dt$HPI) ~ log(dt$Personal_Income) + log(dt$Poverty_Percentage) + log ...



```r
car::vif(m1)
```

```
##    log(dt$Personal_Income) log(dt$Poverty_Percentage)
##                   2.782747                   2.310095
##  log(dt$Unemployment_Rate)          log(dt$Population)
##                   1.416752                   1.449525
##             dt$SomeCollege  log(dt$BachelorAndHigher)
##                   1.094281                   2.637371
```