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:2232        Length:2232        Min.   : 96.64   Min.   : 27415
##  Class :character   Class :character   1st Qu.:139.13   1st Qu.: 38812
##  Mode  :character   Mode  :character   Median :152.77   Median : 43979
##                                        Mean   :160.02   Mean   : 46506
##                                        3rd Qu.:176.78   3rd Qu.: 50729
##                                        Max.   :395.90   Max.   :229825
##  Poverty_Percentage   Population        HighSchoolLess  HighSchoolOnly
##  Min.   : 2.7       Min.   :    1129   Min.   : 1.50   Min.   : 7.80
##  1st Qu.: 9.8       1st Qu.:   19263   1st Qu.: 8.20   1st Qu.:28.90
##  Median :12.7       Median :   37680   Median :11.10   Median :33.70
##  Mean   :13.4       Mean   :  132876   Mean   :12.18   Mean   :33.44
##  3rd Qu.:16.1       3rd Qu.:   97008   3rd Qu.:15.20   3rd Qu.:38.50
##  Max.   :38.2       Max.   :10039107   Max.   :43.10   Max.   :54.50
##   SomeCollege   BachelorAndHigher Unemployment_Rate
##  Min.   :11.2   Min.   : 8.20     Min.   : 1.60
##  1st Qu.:27.9   1st Qu.:16.30     1st Qu.: 3.10
##  Median :31.0   Median :20.90     Median : 3.70
##  Mean   :31.1   Mean   :23.28     Mean   : 3.89
##  3rd Qu.:34.2   3rd Qu.:28.30     3rd Qu.: 4.50
##  Max.   :47.3   Max.   :75.30     Max.   :18.30
```

Correlation Check

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

Education parameters

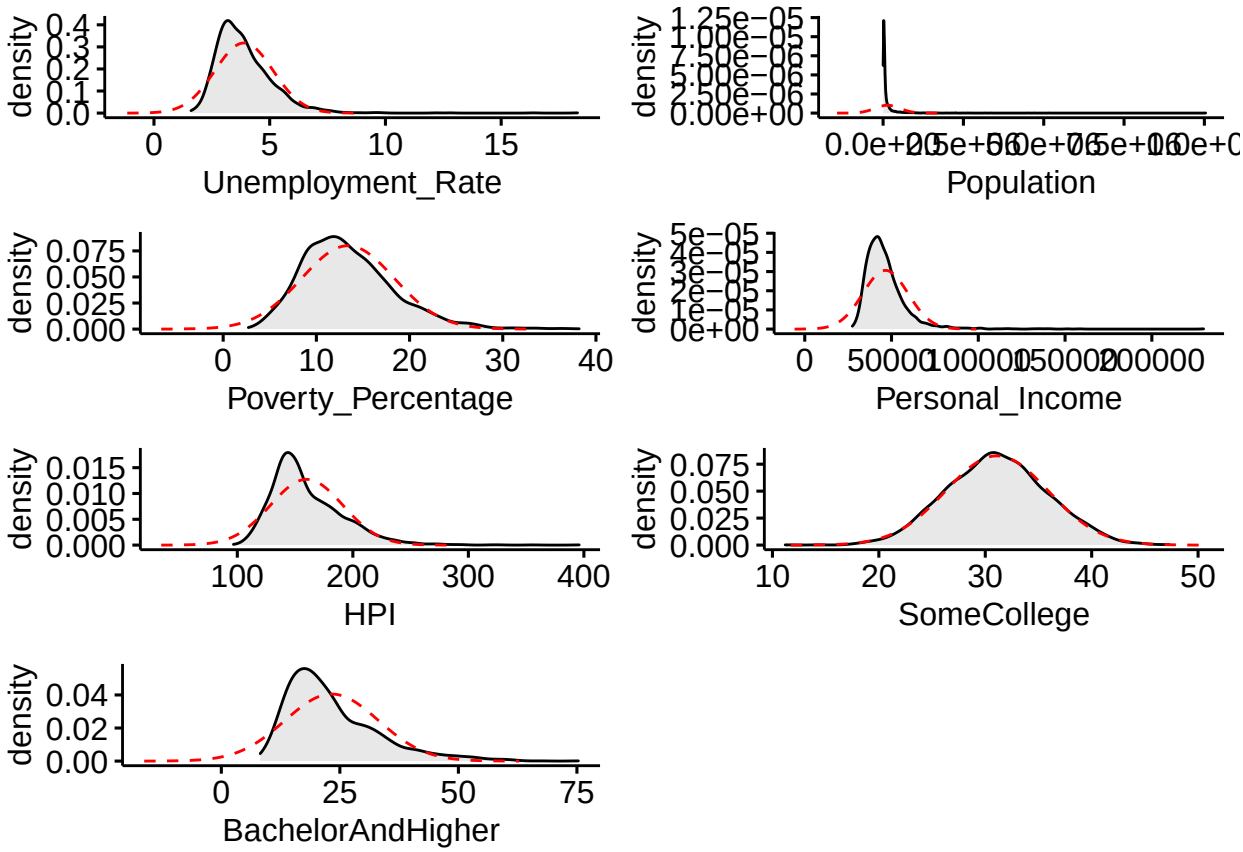
```
##                   HighSchoolLess HighSchoolOnly SomeCollege BachelorAndHigher
## HighSchoolLess         1.0000000      0.2778500  -0.3452977        -0.5976387
## HighSchoolOnly         0.2778500      1.0000000  -0.2031889        -0.7989387
## SomeCollege           -0.3452977     -0.2031889   1.0000000        -0.1458173
## BachelorAndHigher     -0.5976387     -0.7989387  -0.1458173         1.0000000
```

Histogram

```
library(ggpubr)
```

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

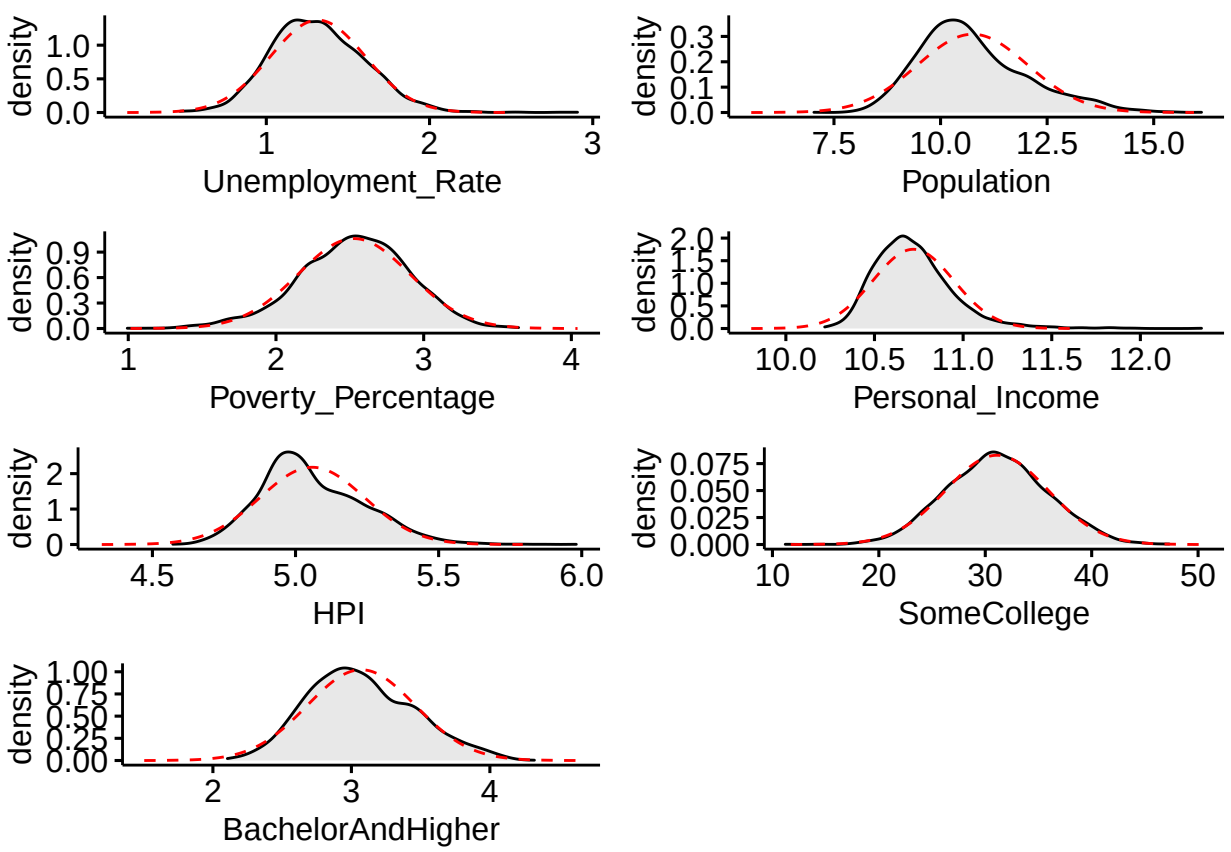
```
a<-ggdensity(dt, x = "Unemployment_Rate", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
b<-ggdensity(dt, x = "Population", fill = "lightgray") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
c<-ggdensity(dt, x = "Poverty_Percentage", fill = "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

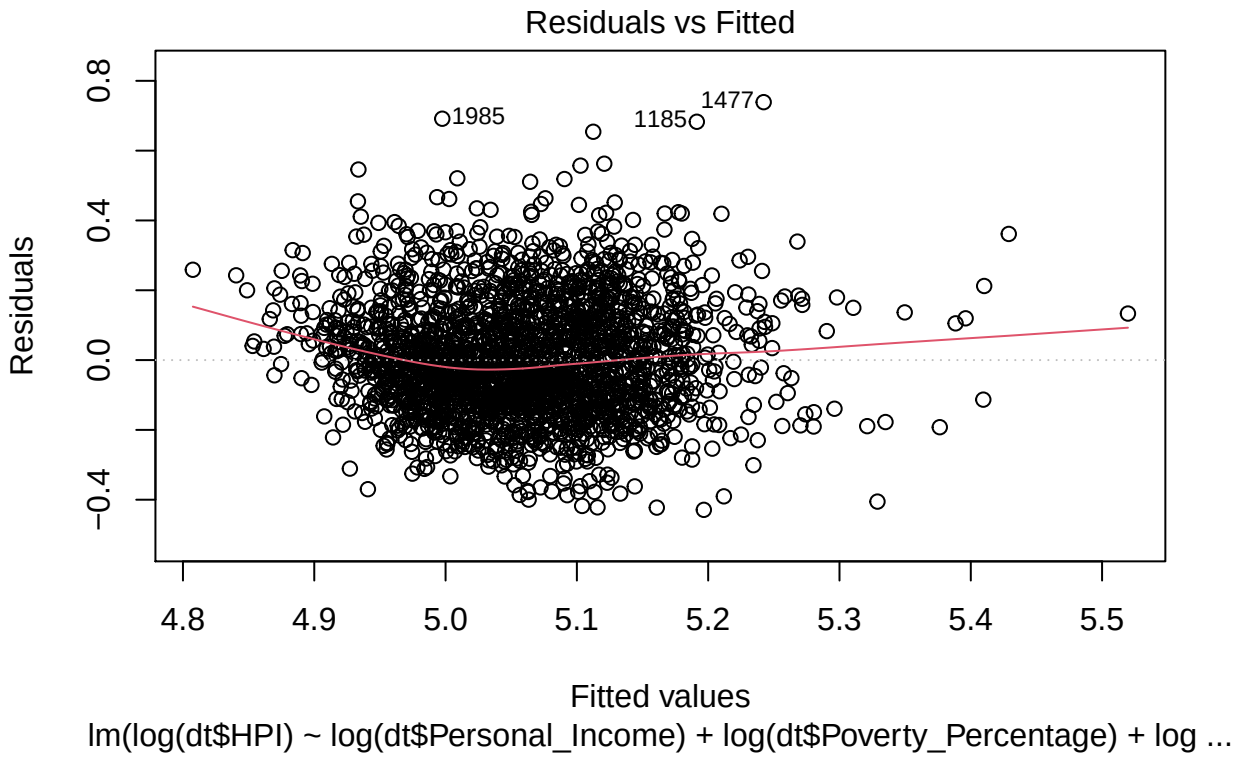
```
m1 = lm(log(dt$HPI)~log(dt$Personal_Income)+log(dt$Poverty_Percentage)+log(dt$Unemployment_Rate)+log(dt$
summary(m1)
```

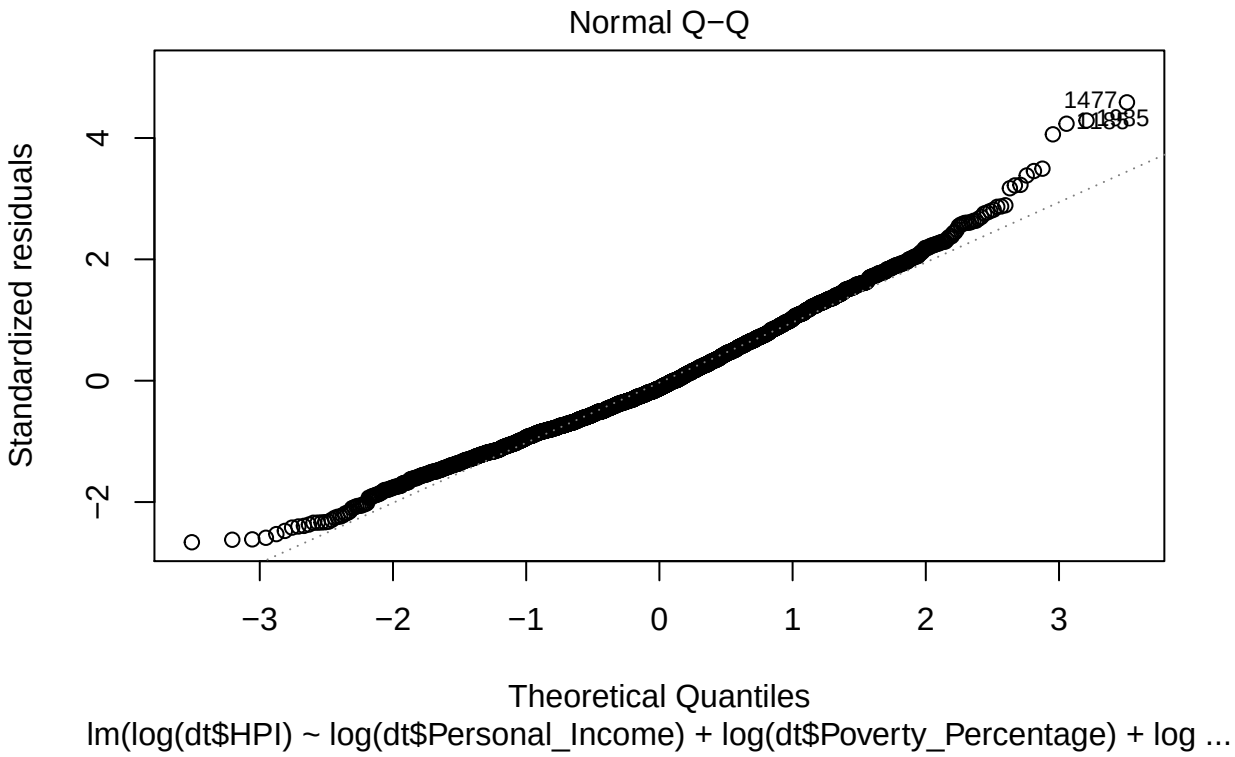
```
##
## Call:
## lm(formula = log(dt$HPI) ~ log(dt$Personal_Income) + log(dt$Poverty_Percentage) +
##     log(dt$Unemployment_Rate) + log(dt$Population) + dt$SomeCollege +
##     log(dt$BachelorAndHigher))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42898 -0.11299 -0.01997  0.10281  0.73889
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.360647   0.280639   4.848 1.33e-06 ***
## log(dt$Personal_Income)       0.304079   0.025521  11.915  < 2e-16 ***
## log(dt$Poverty_Percentage)    0.074468   0.013910   5.353 9.51e-08 ***
## log(dt$Unemployment_Rate)    -0.079504   0.014297  -5.561 3.01e-08 ***
```

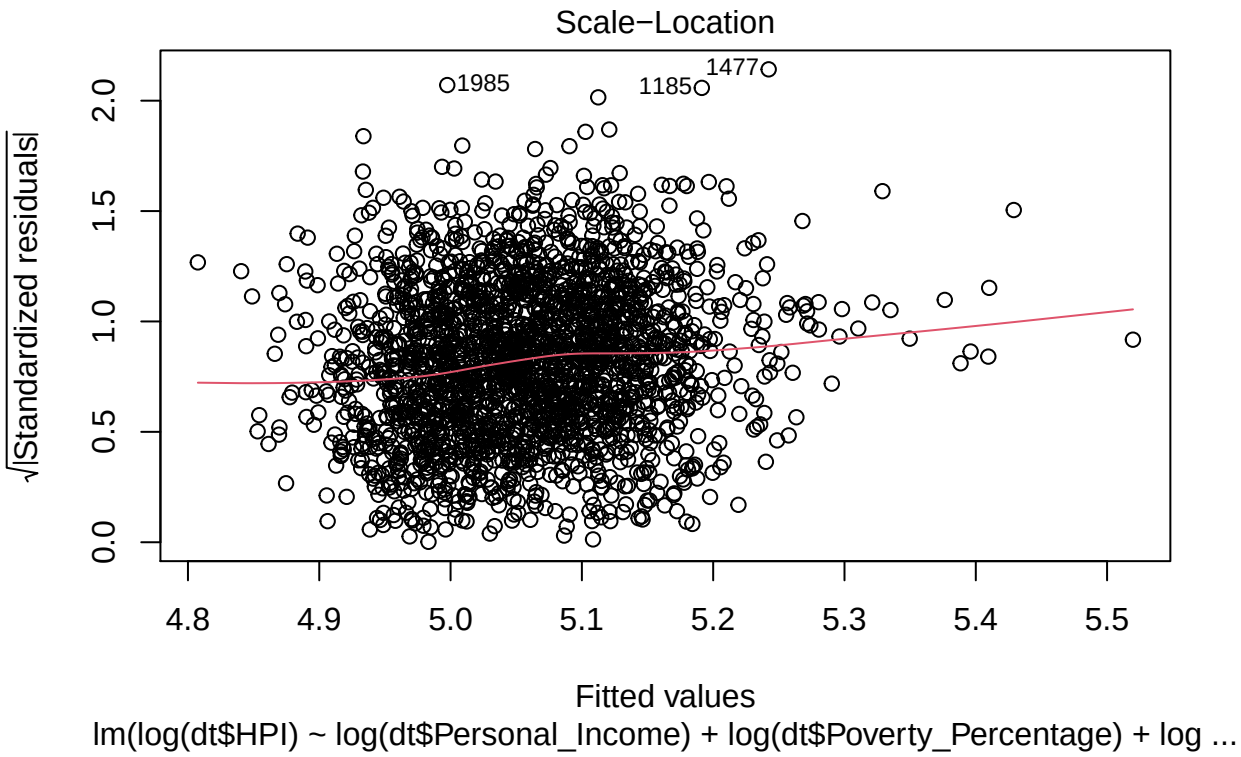
```
## log(dt$Population)          0.004923   0.003244   1.517  0.12932
## dt$SomeCollege              0.005438   0.000754   7.212 7.53e-13 ***
## log(dt$BachelorAndHigher)   0.043153   0.014390   2.999  0.00274 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1616 on 2225 degrees of freedom
## Multiple R-squared:  0.1953, Adjusted R-squared:  0.1931
## F-statistic: 89.98 on 6 and 2225 DF,  p-value: < 2.2e-16
```

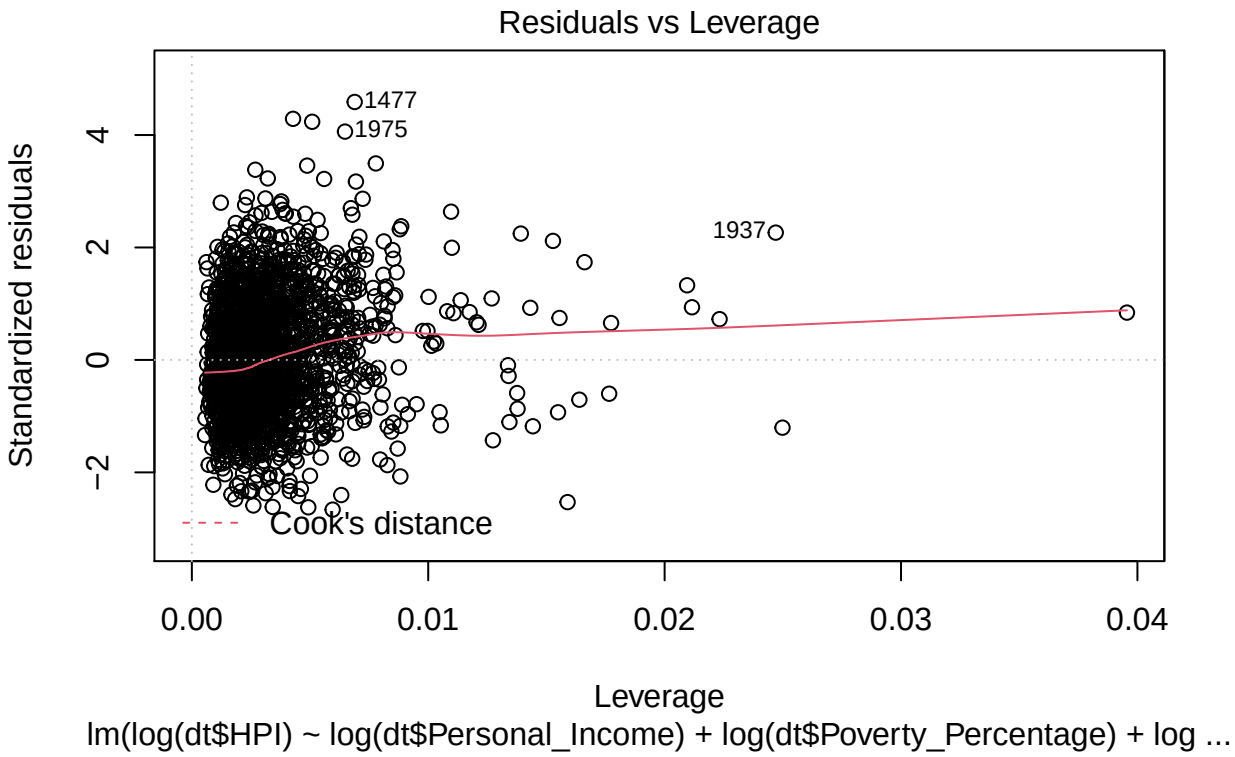
Diagnostic Plots

```
plot(m1)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(dt$HPI) ~ log(dt$Personal_Income) + log(dt$Poverty_Percentage) + log ...
1477
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
1985
1185
1477
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
4.8
4.9
5.0
5.1
5.2
5.3
5.4
5.5
Fitted values
lm(log(dt$HPI) ~ log(dt$Personal_Income) + log(dt$Poverty_Percentage) + log ...


Residuals vs Leverage
Standardized residuals
4
2
0
−2
1477
1975
1937
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(log(dt$HPI) ~ log(dt$Personal_Income) + log(dt$Poverty_Percentage) + log ...