NL_Modeling

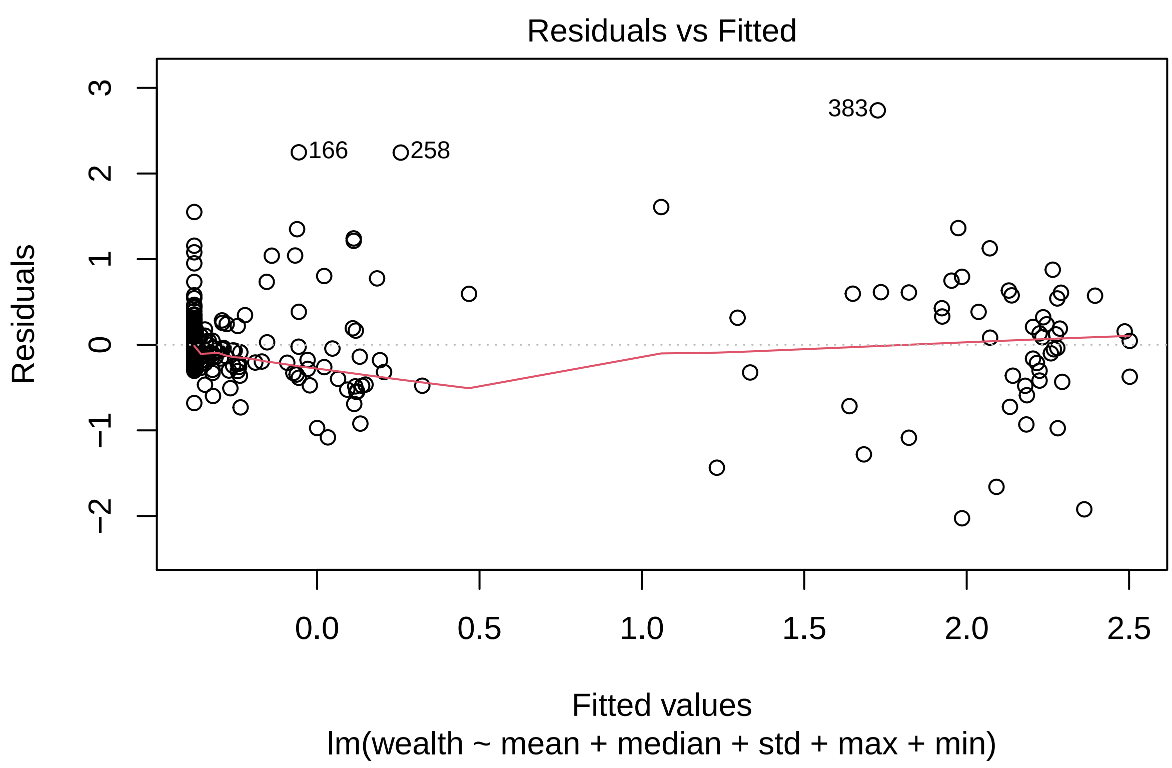
Analysis of Night Light Data and Poverty

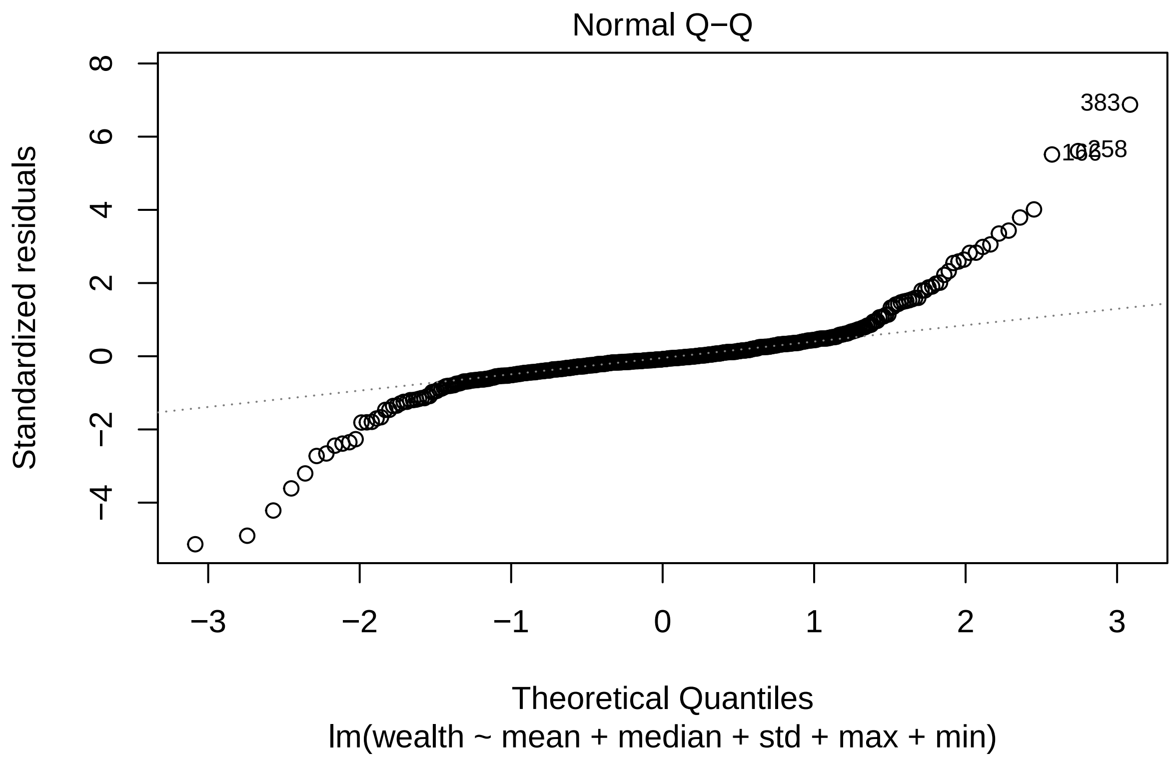
Loading main dataframe:

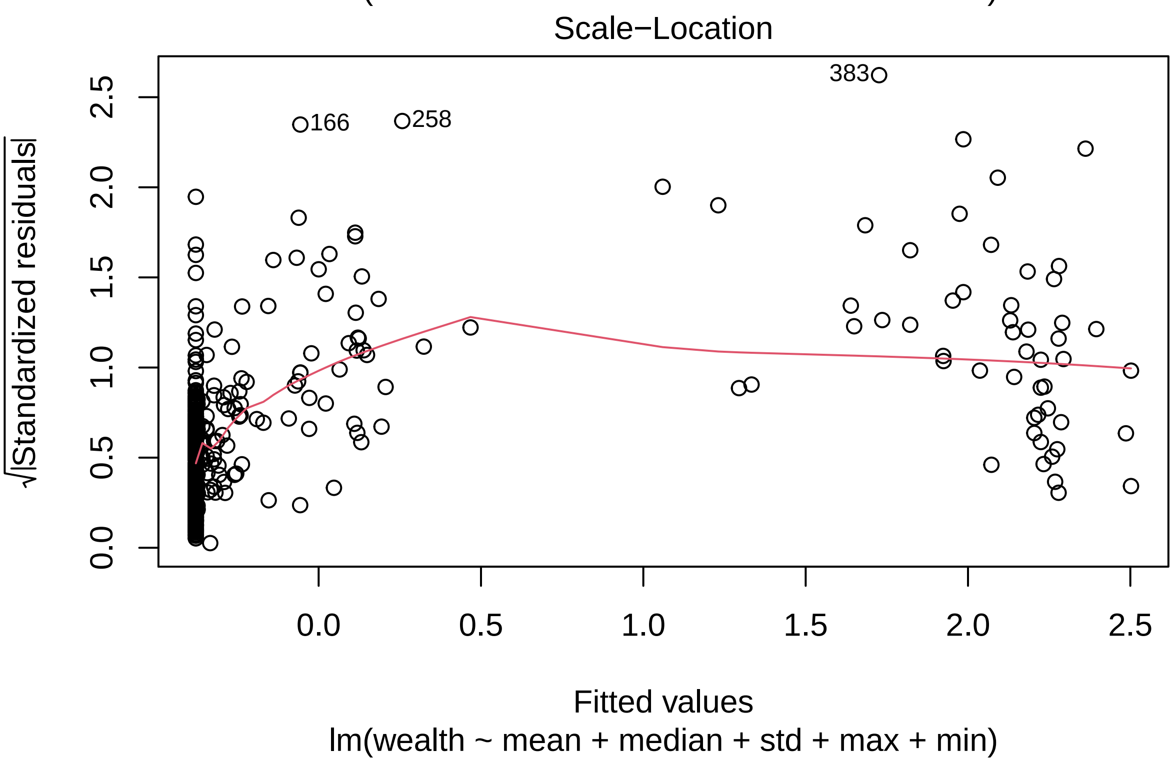
```
df <- read.csv("../processed/nl_wealth_data.csv")
```

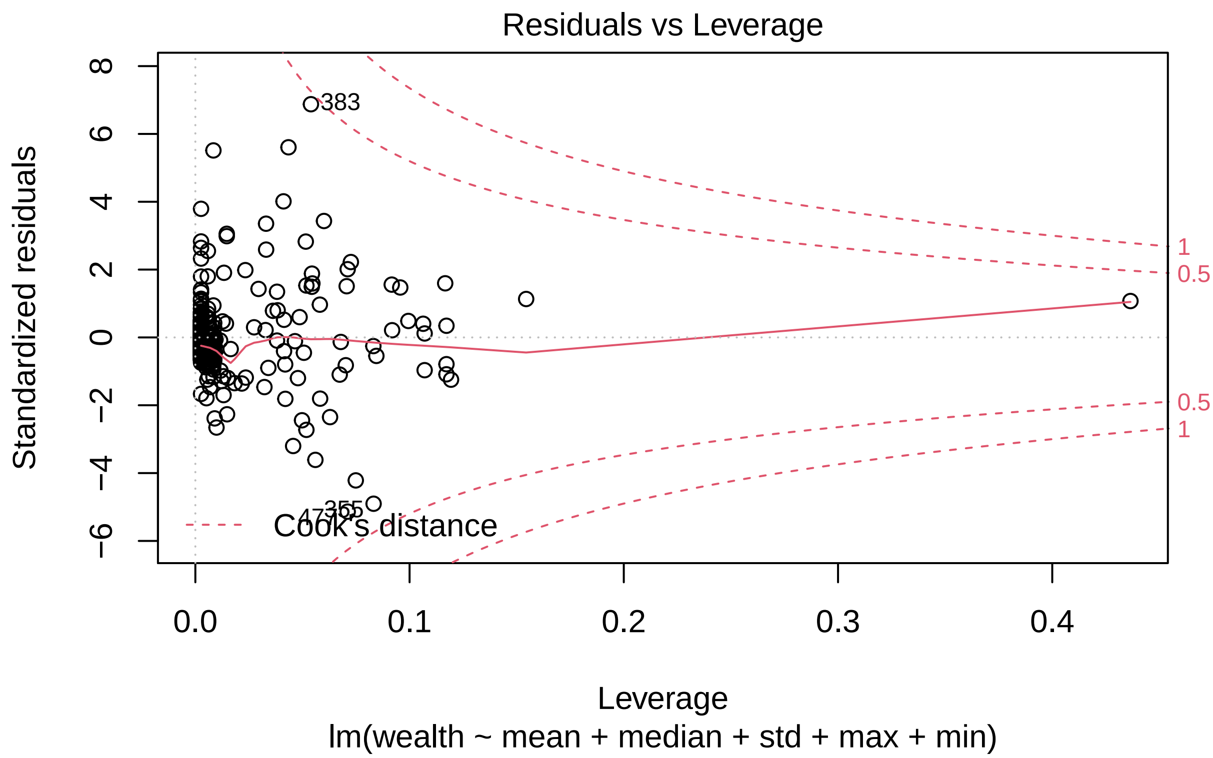
Simple Linear Model of all variables

```
lm <- lm(wealth~mean + median + std + max + min, data=df)
plot(lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(wealth ~ mean + median + std + max + min)
383
166
258


Scale−Location
√|Standardized residuals|
Fitted values
lm(wealth ~ mean + median + std + max + min)
383
166
258




```r
summary(lm)
```

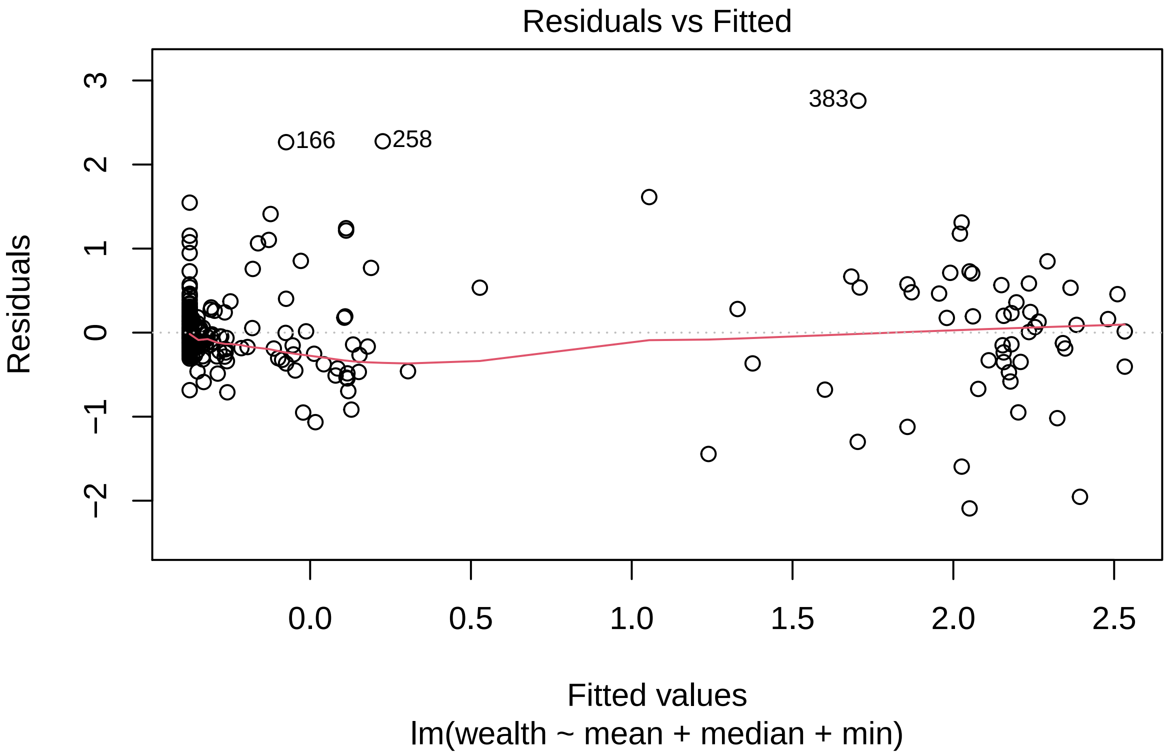
```
##
## Call:
## lm(formula = wealth ~ mean + median + std + max + min, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02654 -0.14144 -0.03232  0.10472  2.73770
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.378228   0.021127 -17.902  < 2e-16 ***
## mean         0.110431   0.030473   3.624 0.000321 ***
## median      -0.037731   0.019537  -1.931 0.054028 .
## std          0.033970   0.032773   1.037 0.300477
## max         -0.003616   0.008410  -0.430 0.667438
## min         -0.064462   0.028274  -2.280 0.023043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4094 on 486 degrees of freedom
## Multiple R-squared:  0.7737, Adjusted R-squared:  0.7714
## F-statistic: 332.4 on 5 and 486 DF,  p-value: < 2.2e-16
```

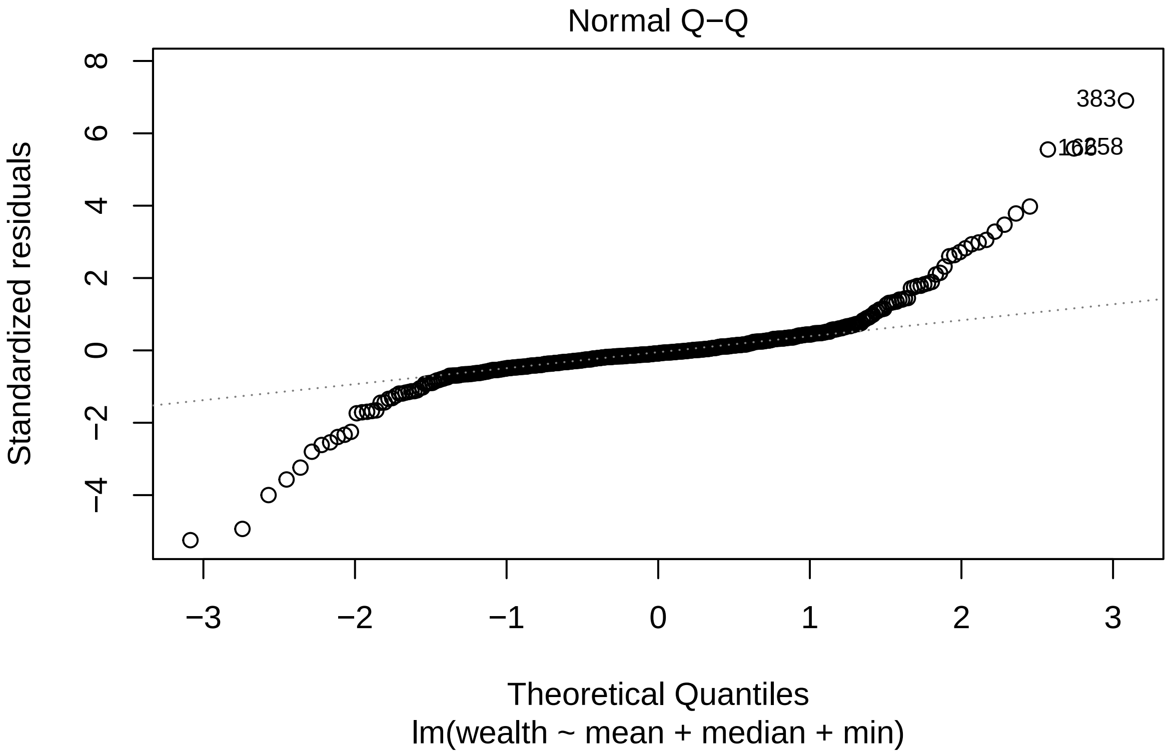
Observe that the quantile plot shows severe deviation from normality and outliers. This can be potentially resolved with bootstrapping. Further, I beleive that most of the outliers correspond to cities and therefore have a much larger amount of nightlight than surrounding regions but potentially not as much wealth.

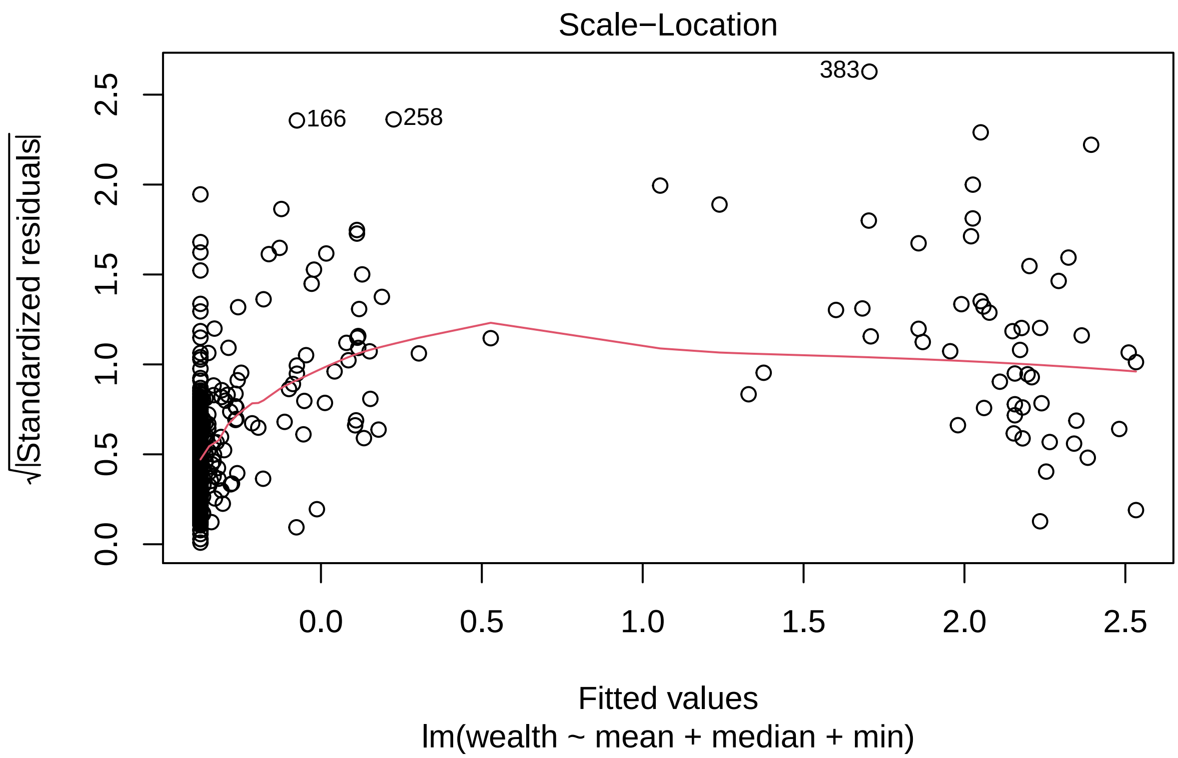
Further, note that STD, MAX, MEDIAN are not significant at the $\alpha = 0.01$ level. We remove them from the model in later steps.

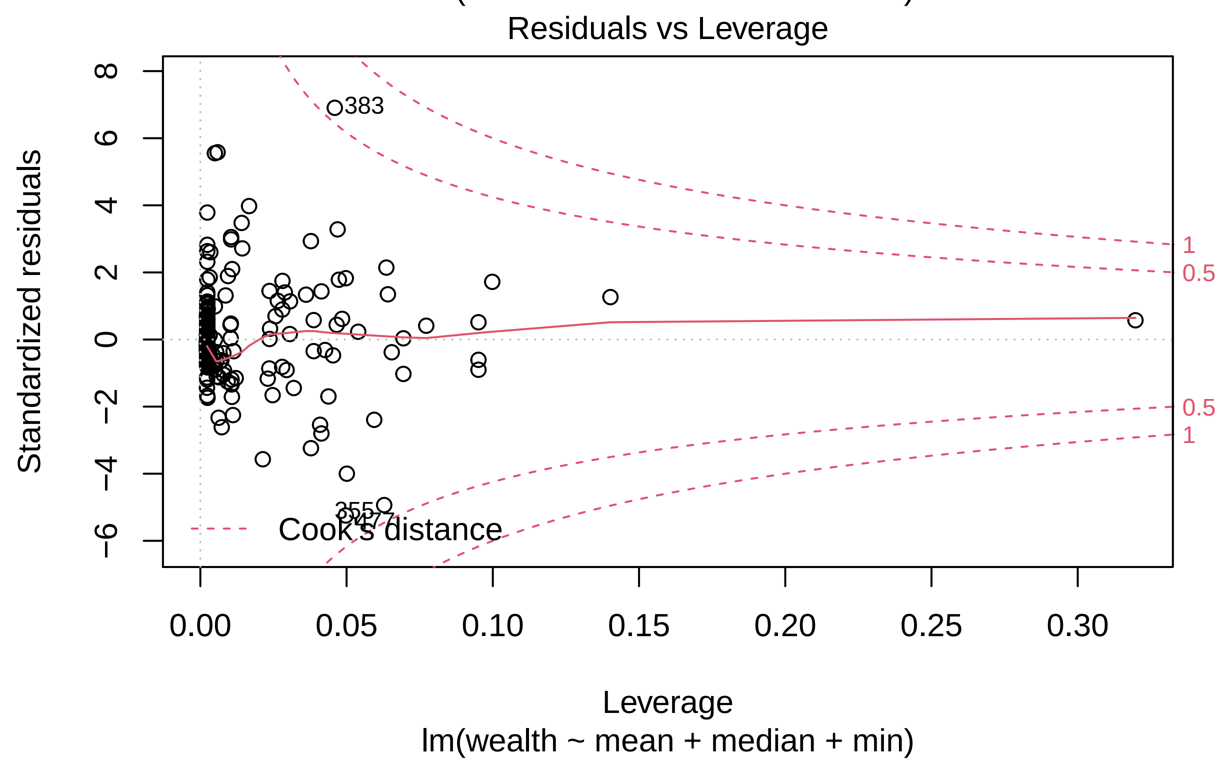
Simple Linear Model of only mean, median, min

```r
lm <- lm(wealth~mean + median + min, data=df)
plot(lm)
```










```
summary(lm)
```

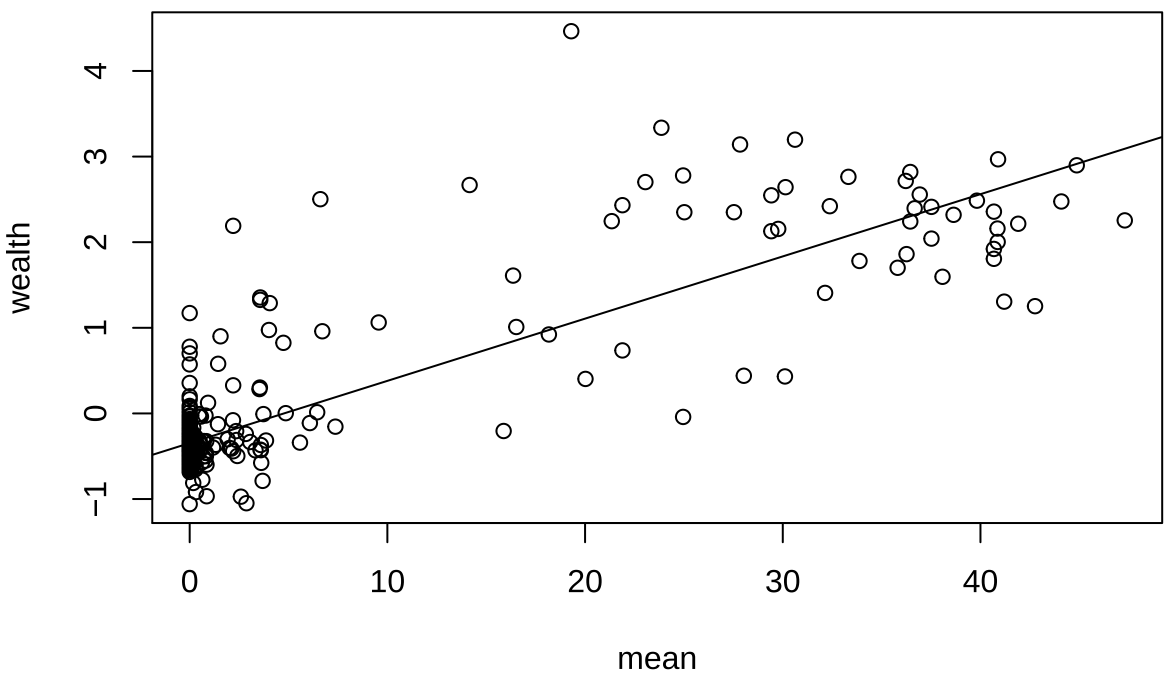
```
##
## Call:
## lm(formula = wealth ~ mean + median + min, data = df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09146 -0.14174 -0.03392  0.10134  2.75922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.37453    0.01997 -18.751  < 2e-16 ***
## mean         0.13625    0.01161  11.738  < 2e-16 ***
## median      -0.05004    0.01424  -3.515  0.00048 ***
## min         -0.07933    0.02407  -3.297  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.409 on 488 degrees of freedom
## Multiple R-squared:  0.7731, Adjusted R-squared:  0.7717
## F-statistic: 554.4 on 3 and 488 DF,  p-value: < 2.2e-16
```

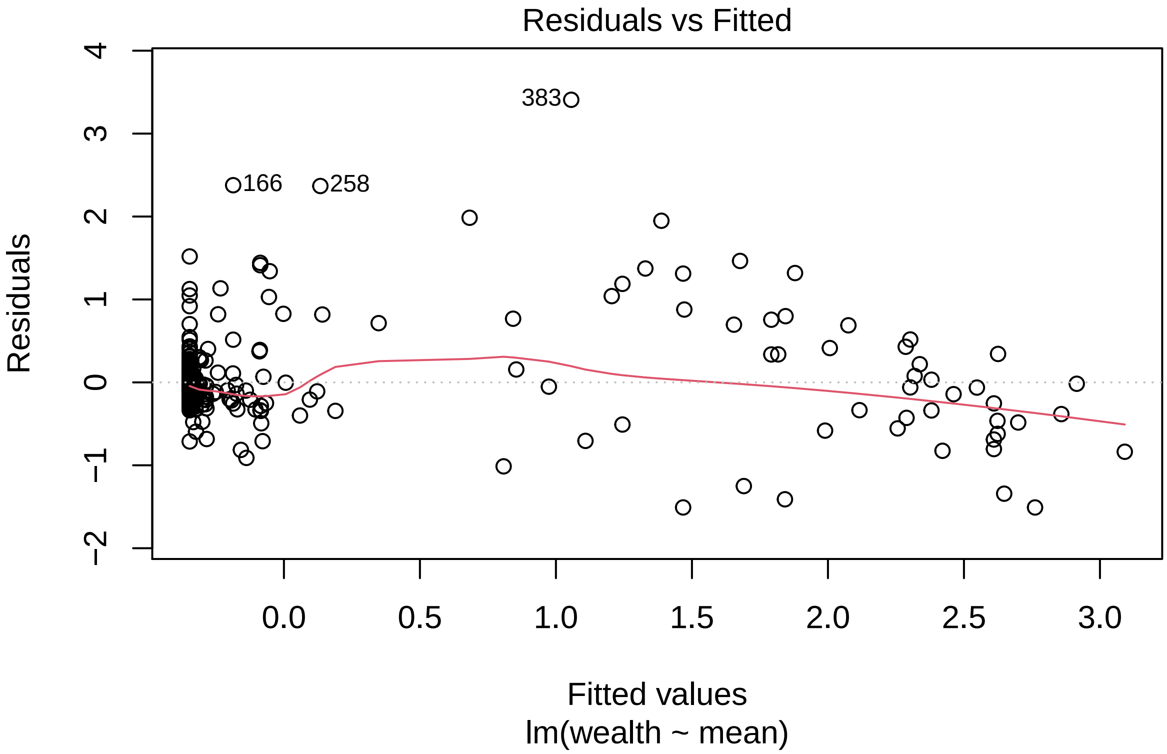
This model seems to have a very high r^2_{adj} value. However, it still shows deviations from normality and outliers.

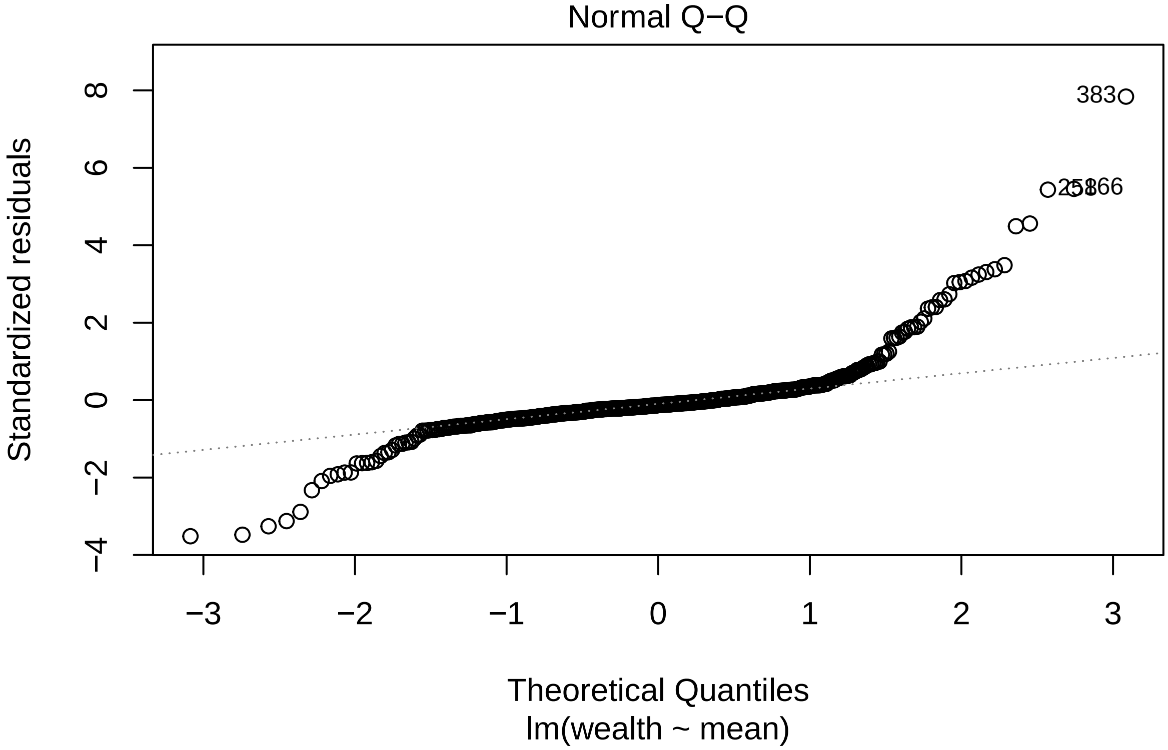
Simple Linear Model of only mean

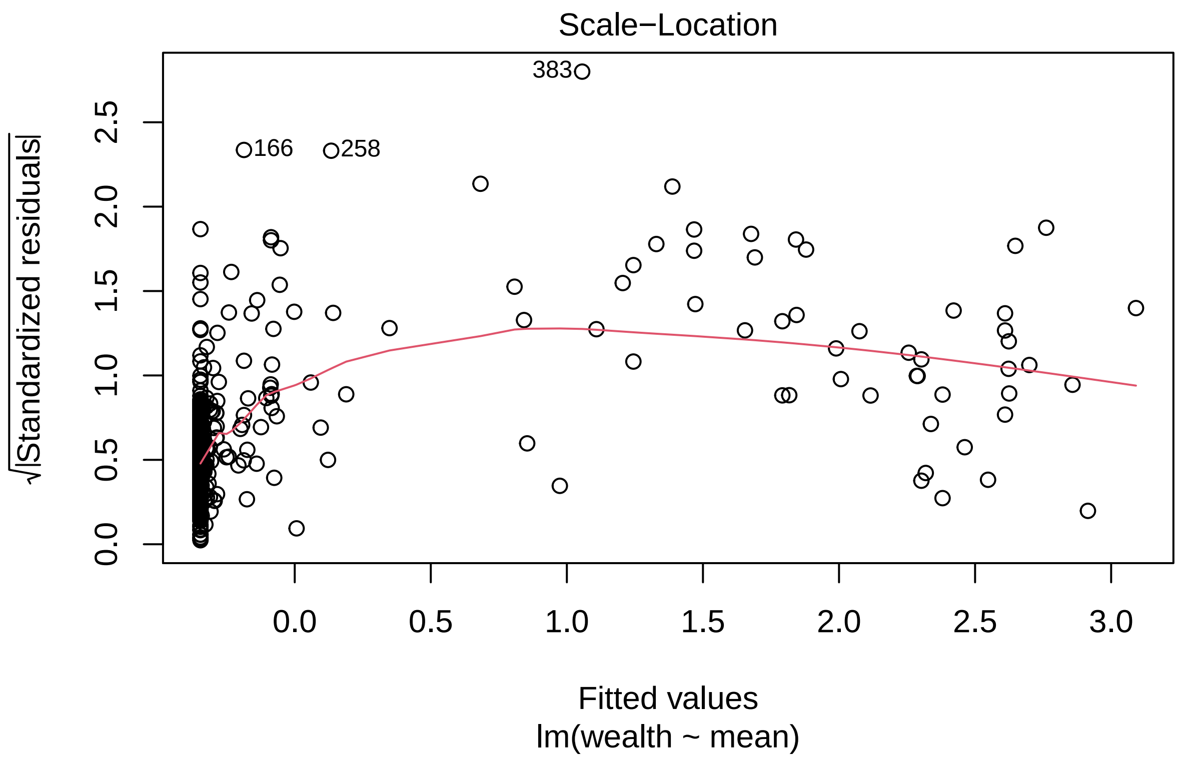
```r
lm <- lm(wealth~mean, data=df)
plot(wealth~mean, data=df)
abline(lm)
```

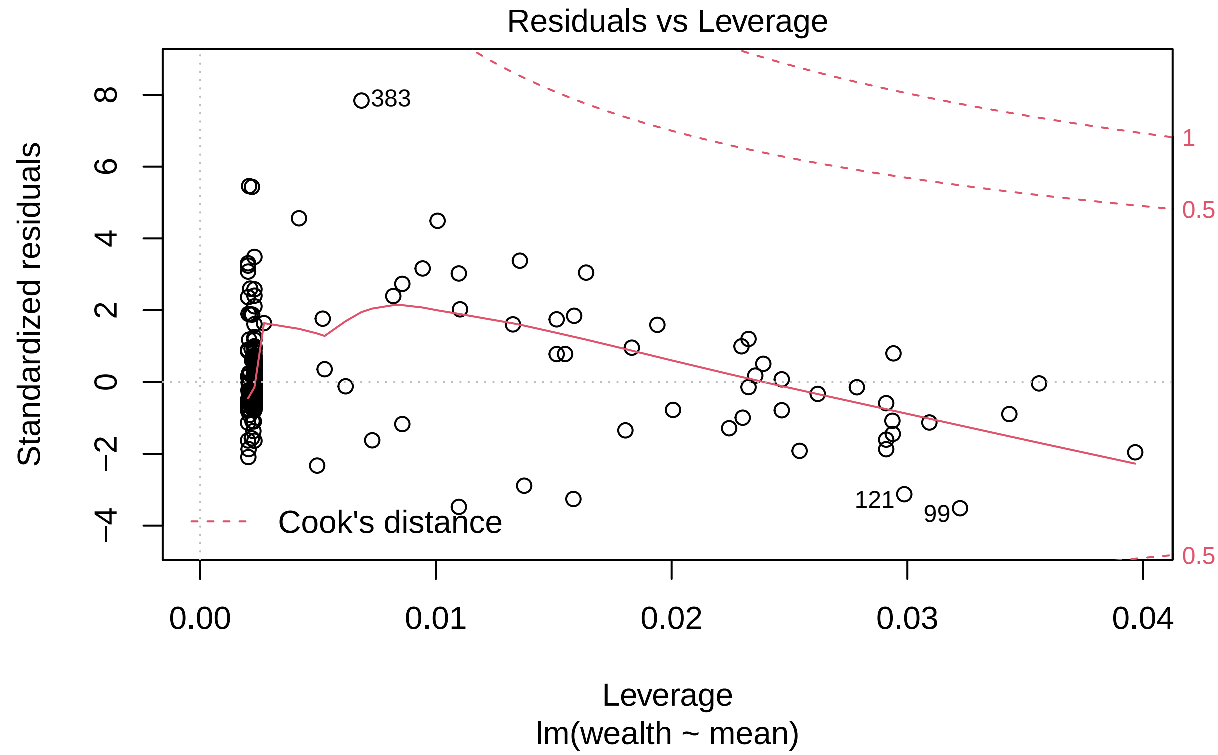



```r
plot(lm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(wealth ~ mean)
383
166
258
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−2
−1
0
1
2
3
4


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(wealth ~ mean)
383
258
166
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
6
8






```
summary(lm)
```

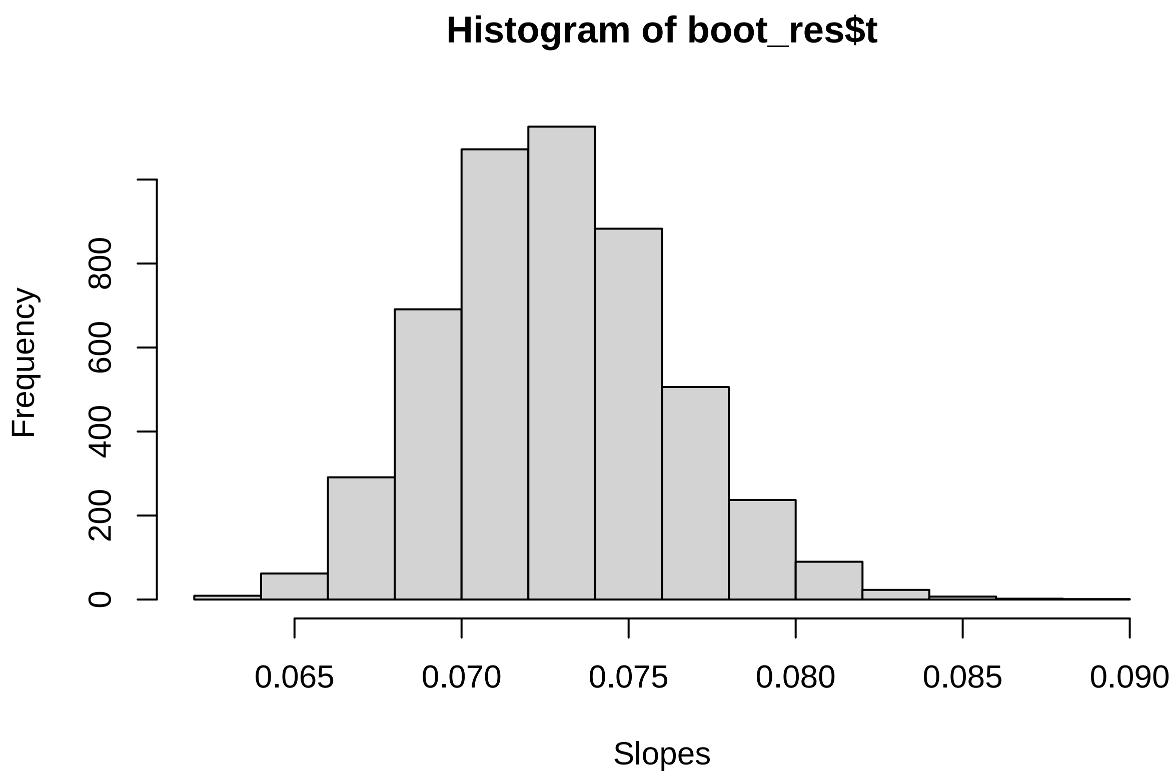
```
##
## Call:
## lm(formula = wealth ~ mean, data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5087 -0.1593 -0.0560  0.0733  3.4077
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.346517   0.020944  -16.55   <2e-16 ***
## mean         0.072681   0.001941   37.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4362 on 490 degrees of freedom
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.7404
## F-statistic:  1402 on 1 and 490 DF,  p-value: < 2.2e-16
```

Here is just a plot of mean. Also, you have a chance to actually see the data and the strong correlation.

Bootstrapped Linear Model only mean

```
set.seed(42)
N <- 5000
boot_func <- function(d, i) {
    return(summary(lm(wealth~mean, data=d[i,]))$coeff[2])
}
boot_res <- boot(df, boot_func, R = N)
hist(boot_res$t, xlab="Slopes")
```




```
boot.ci(boot_res)
```

```
## Warning in boot.ci(boot_res): bootstrap variances needed for studentized
## intervals
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.0658,  0.0793 )   ( 0.0654,  0.0787 )
##
## Level     Percentile            BCa
## 95%   ( 0.0666,  0.0800 )   ( 0.0666,  0.0800 )
## Calculations and Intervals on Original Scale
```

Bootstrapping gives the confidence interval for the slope. I need to investigate it more thoroughly.