# Crime Data Analysis

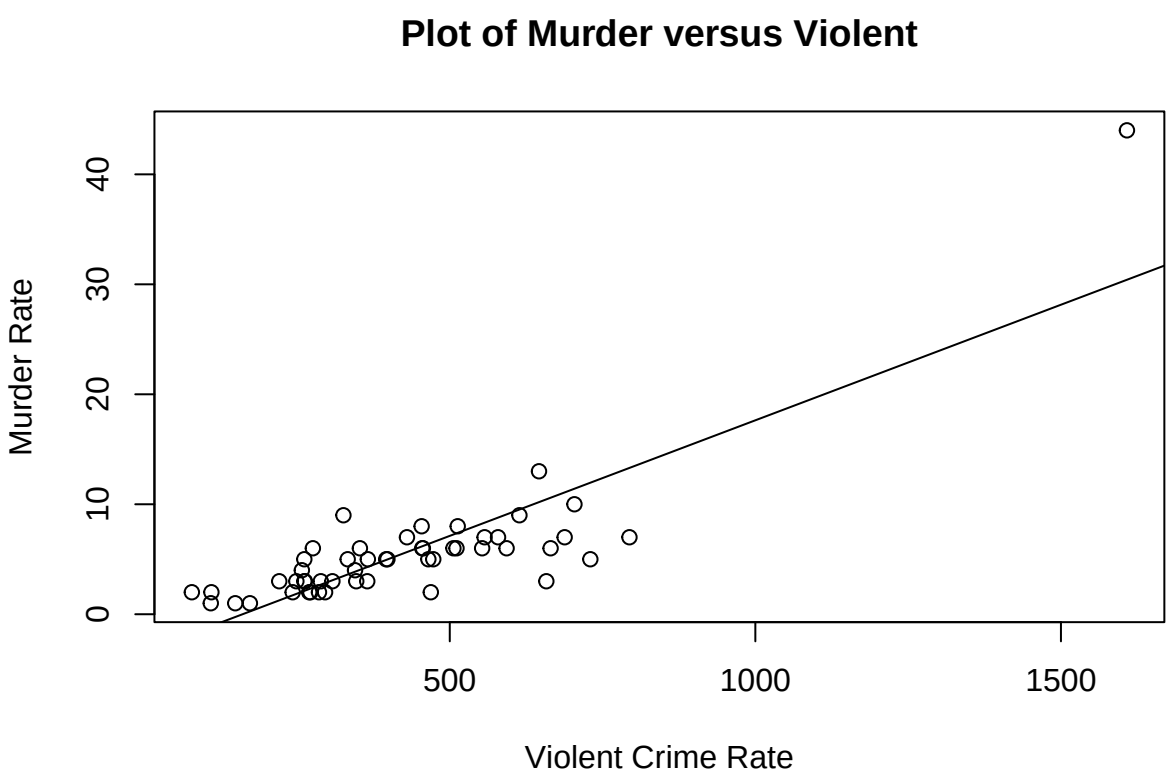
Bonald So

The data originate from the Statistical Abstract of the United States for the year 2005, and were downloaded from Alan Agresti's site at the University of Florida: https://users.stat.ufl.edu/~aa/smss/data/Crime2.dat

```
load(file.path("UScrime05.Rda"))
fit1 = lm(murder~violent, data = UScrime05)
summary(fit1)
```

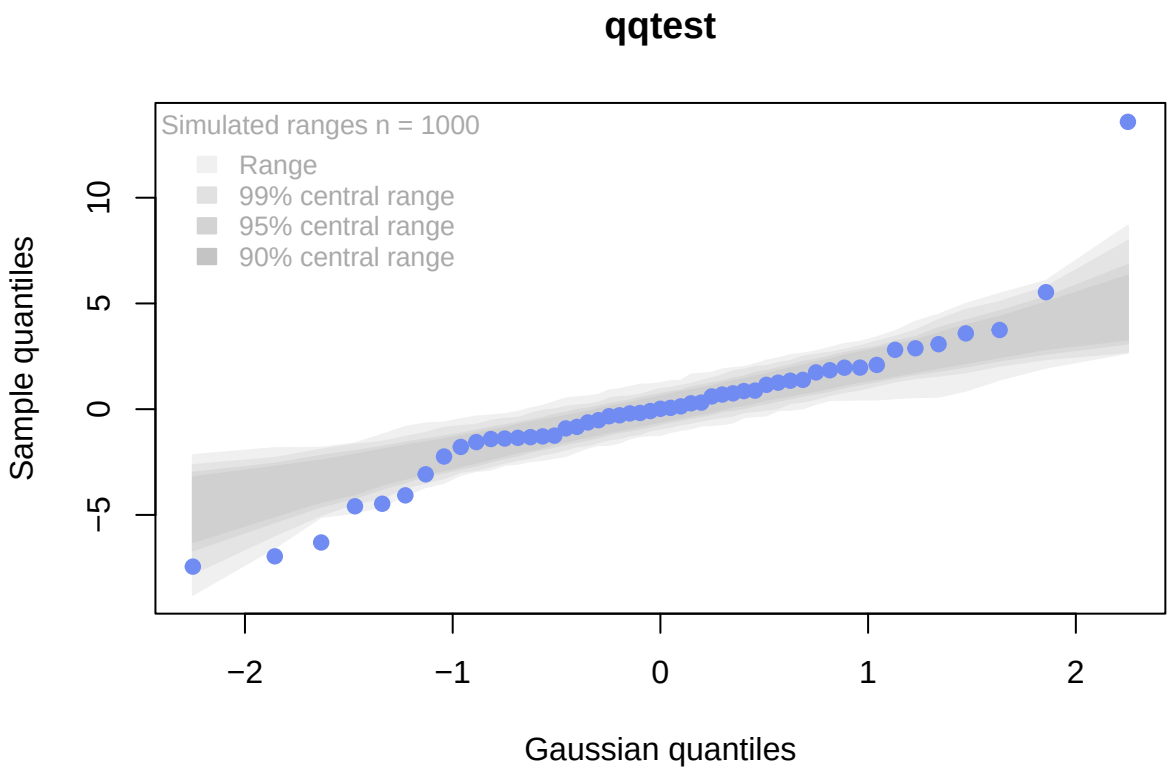
```
##
## Call:
## lm(formula = murder ~ violent, data = UScrime05)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4457 -1.3394  0.0191  1.3659 13.5869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.384359   0.938042  -3.608 0.000723 ***
## violent      0.021018   0.001918  10.957 8.88e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.291 on 49 degrees of freedom
## Multiple R-squared:  0.7102, Adjusted R-squared:  0.7043
## F-statistic: 120.1 on 1 and 49 DF,  p-value: 8.876e-15
```

```
plot(UScrime05$violent, UScrime05$murder,
     xlab="Violent Crime Rate", ylab="Murder Rate",
     main="Plot of Murder versus Violent")
abline(fit1)
```



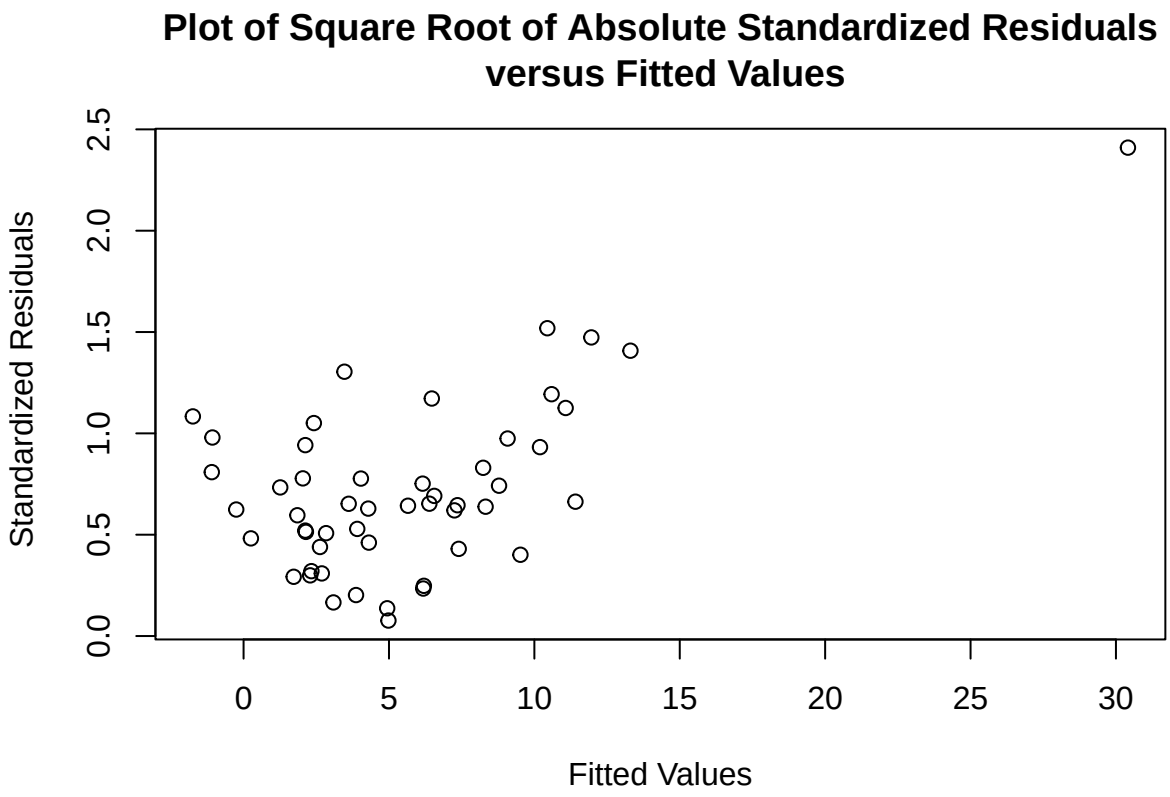

Analysis: From the summary, p-value is less than 0.05. So there is significant difference between violent and murder. So the plot shows increasing mean murder rate of 0.02 under per unit change of violent crime rate. One notable feature from the plot is that there is a data point on the top right which is far away from the majority of the data.

```
library(qqtest)
qqtest(fit1$residuals)
```




From the qqtest, we can see that there are some points lying on the edge of the shaded region, and the top right sample even lies outside the shaded region. So there is strong eveidence against the assumption of normality of the residual distribution.

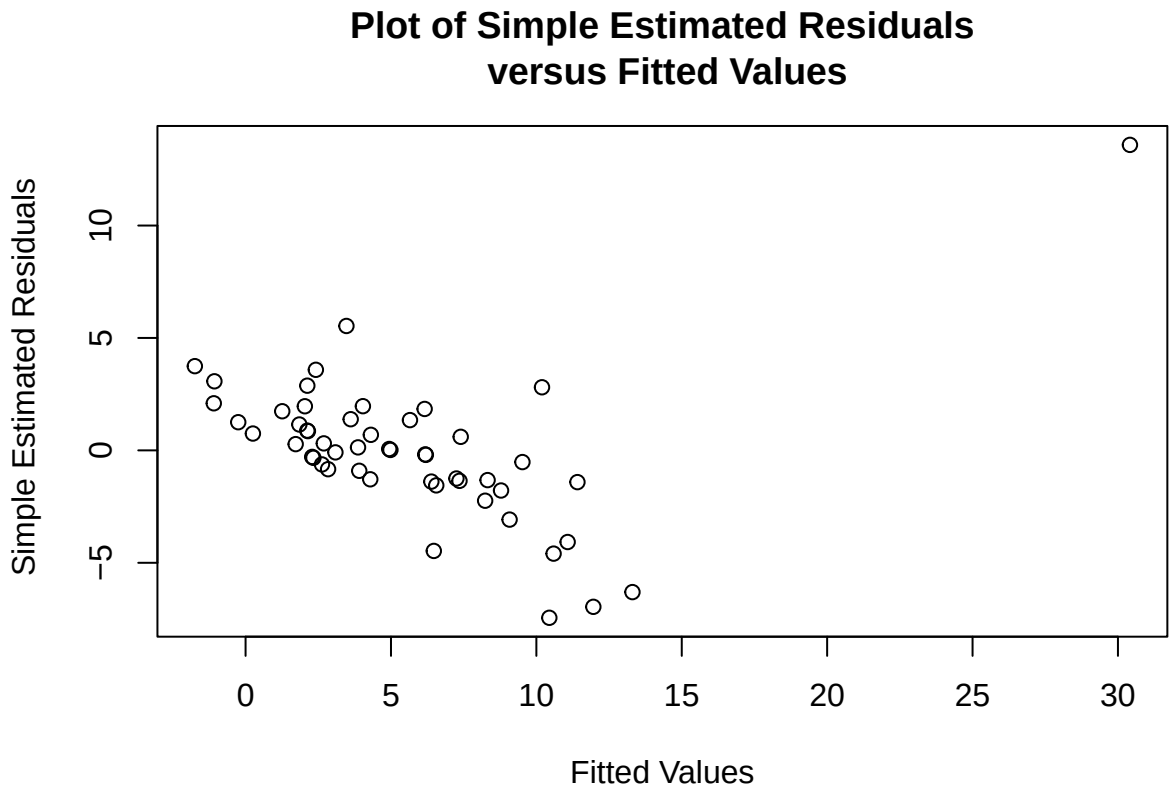
```r
stres = sqrt(abs(rstandard(fit1)))
muhats = fitted(fit1)
plot(muhats, stres, xlab="Fitted Values",
     ylab=" Standardized Residuals",
     main="Plot of Square Root of Absolute Standardized Residuals\n versus Fitted Values")
```




We can see from the plot that there is one data point appearing on the top right which is far away from the majority of the data. Since there is a residual point having square-rooted absolute standard residual greater than 2, we conclude there is evidence against the hypothesis of homoscedasticity.

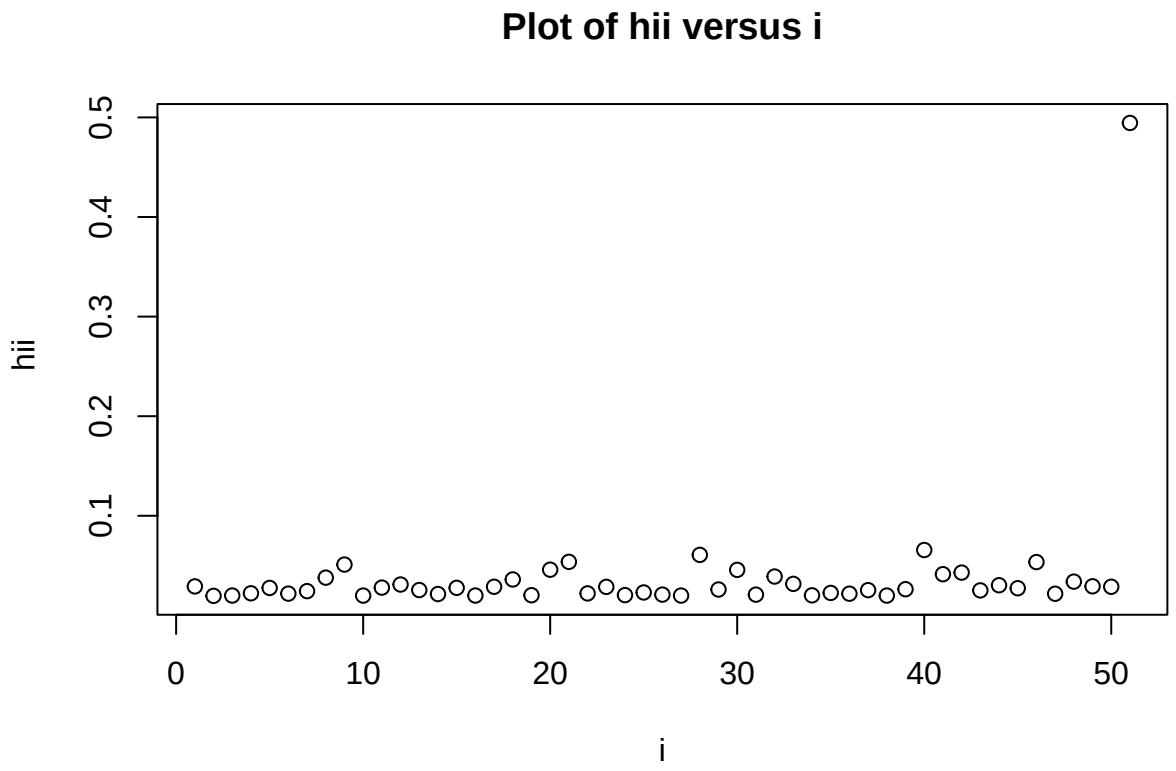
Thought: Plotting versus fitted values might be a good idea to test homoscedasticity because we can see which points are having larger residuals easily.

```r
plot(muhats, fit1$residuals, xlab="Fitted Values",
     ylab=" Simple Estimated Residuals",
     main="Plot of Simple Estimated Residuals\n versus Fitted Values")
```




Simple estimated residuals are used to see if there appears any special pattern to check linearity and equal variance assumption. Excluding the point on the top right corner, it is showing a decreasing trend of residuals with increasing fitted value.

```r
hv = hatvalues(fit1)
xseq = seq(1, 51, length.out = 51)
plot(xseq, hv, xlab="i",
     ylab=" hii",
     main="Plot of hii versus i")
```

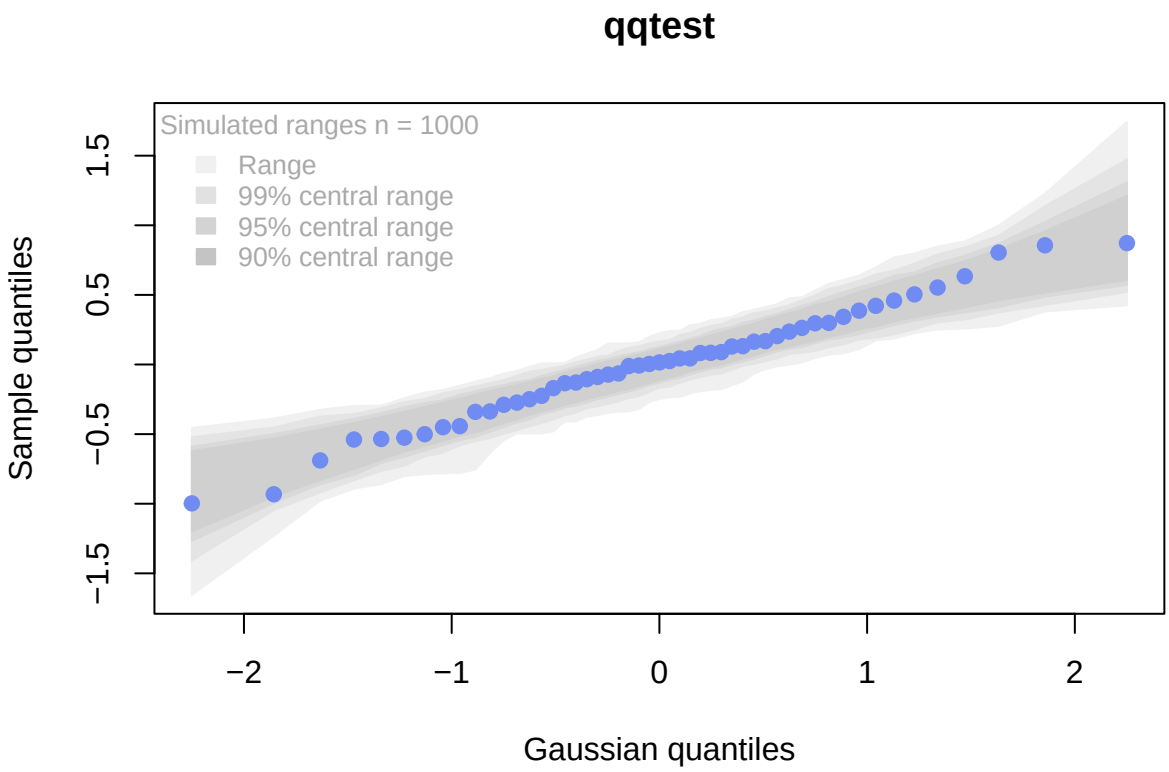



From the plot, we can see that the greatest potential to affect the fitted model would be District of Columbia(51).

```r
fit2 = lm(log(murder)~log(violent), data = UScrime05)
plot(log(UScrime05$violent), log(UScrime05$murder),
     xlab="Log Violent Crime Rate", ylab="Log Murder Rate",
     main="Plot of Log Murder versus Log Violent")
abline(fit2)
```

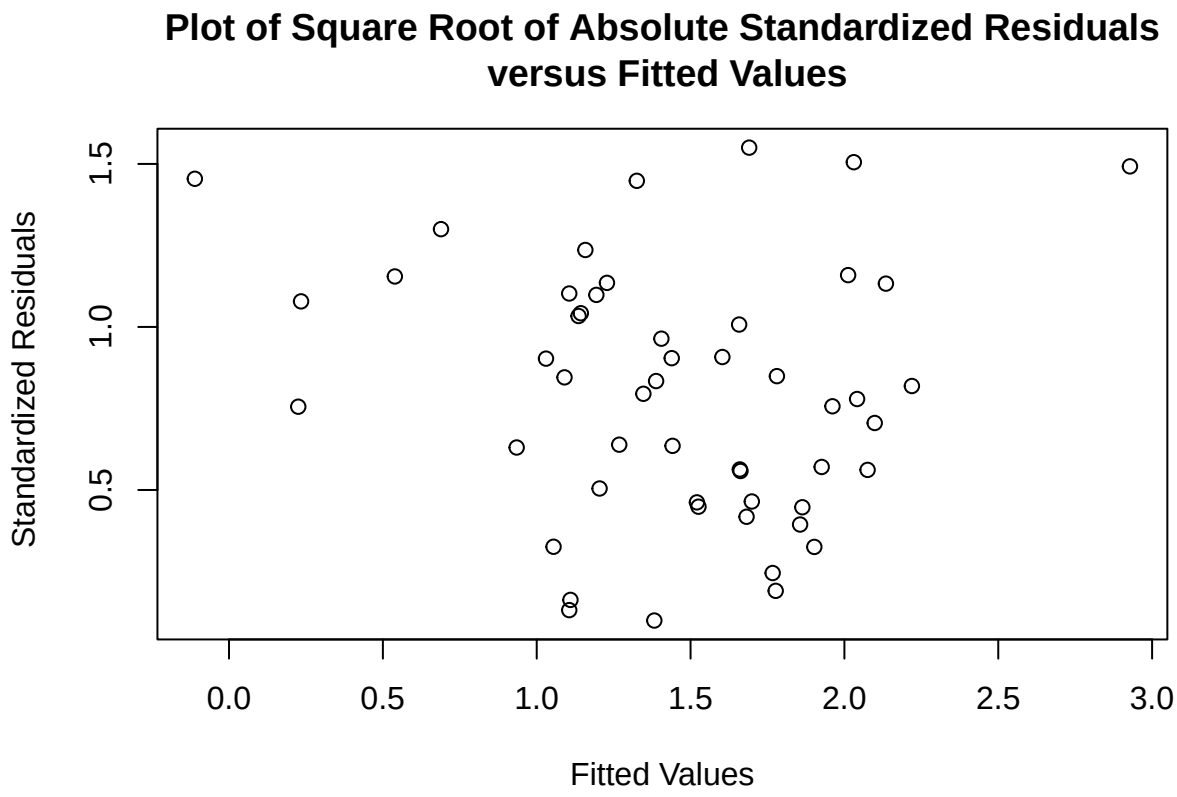

Analysis: We can see from the plot that the mean log(murder) rate increases as log(violent) crime rate increases. Most of data points are located in the center of the graph, while some on the corners but still close to the fitted line.

```
qqtest(fit2$residuals)
```



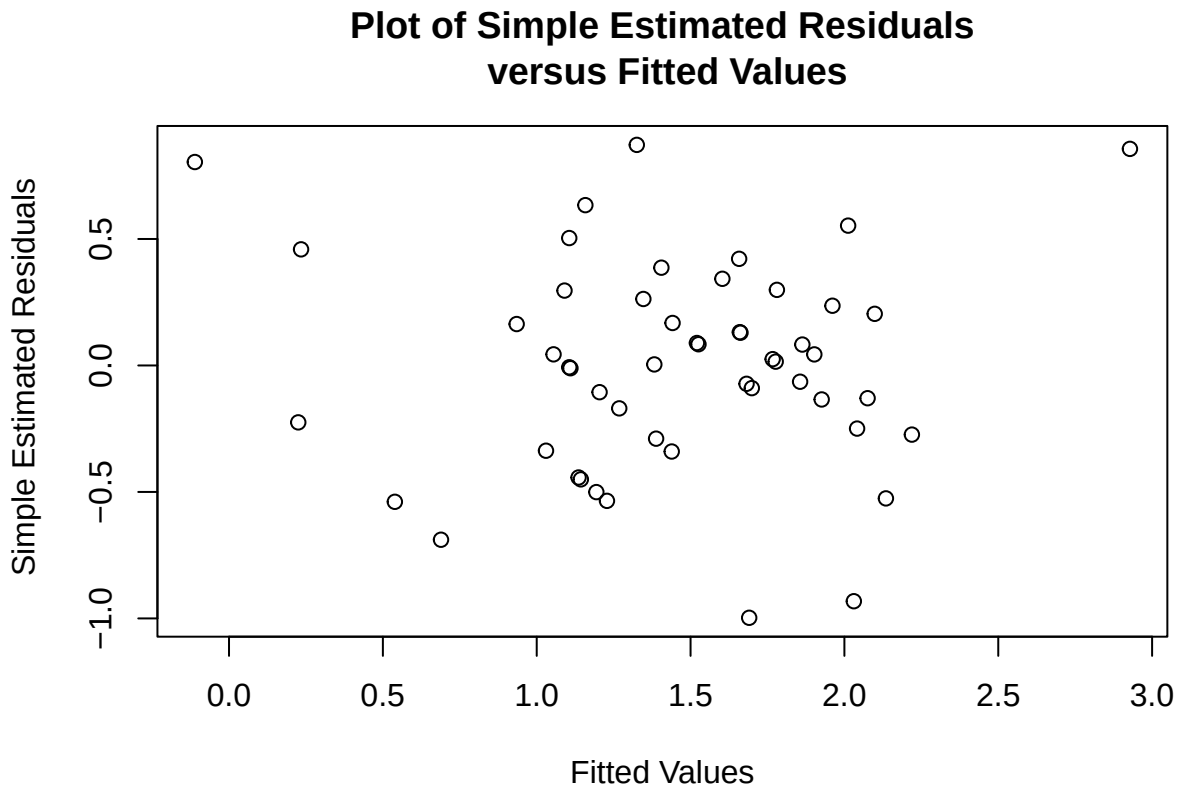

We can see from the plot that all data are lying inside the shaded region. So we conclude that data are normally distributed.

```r
stres2 = sqrt(abs(rstandard(fit2)))
muhats2 = fitted(fit2)
plot(muhats2, stres2, xlab="Fitted Values",
     ylab=" Standardized Residuals",
     main="Plot of Square Root of Absolute Standardized Residuals\n versus Fitted Values")
```



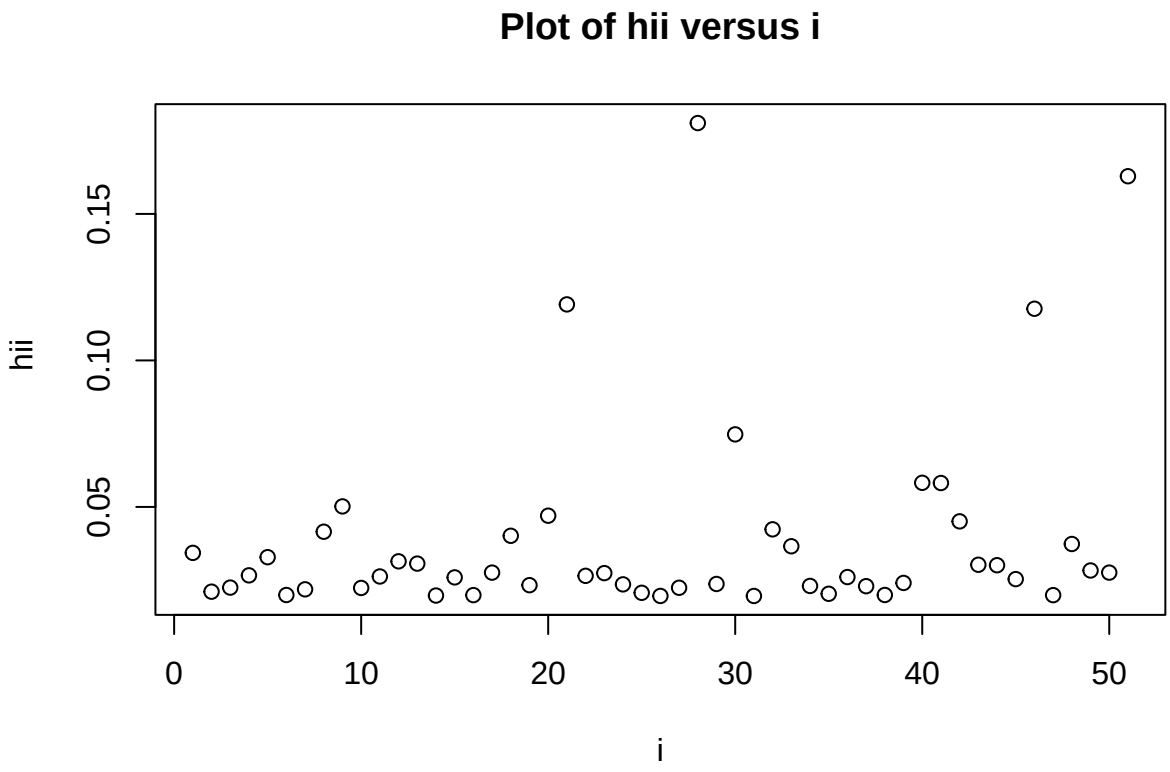

We can see from the plot that all residuals are randomly distributed and within 1.5, so it proves heteroscedasticity increased for fit2 compared to fit1.

```r
plot(muhats2, fit2$residuals, xlab="Fitted Values",
     ylab=" Simple Estimated Residuals",
     main="Plot of Simple Estimated Residuals\n versus Fitted Values")
```




From the plot, we can see data are randomly distributed, and so the structure of decreasing residuals values for increasing fitted values is missing.

```r
hv2 = hatvalues(fit2)
plot(xseq, hv2, xlab="i",
     ylab=" hii",
     main="Plot of hii versus i")
```




From the plot, we can see that the potential influence of District of Columbia has changed after applying log transformation. Beacuse its influence is lowered by other states which also have large hii values.

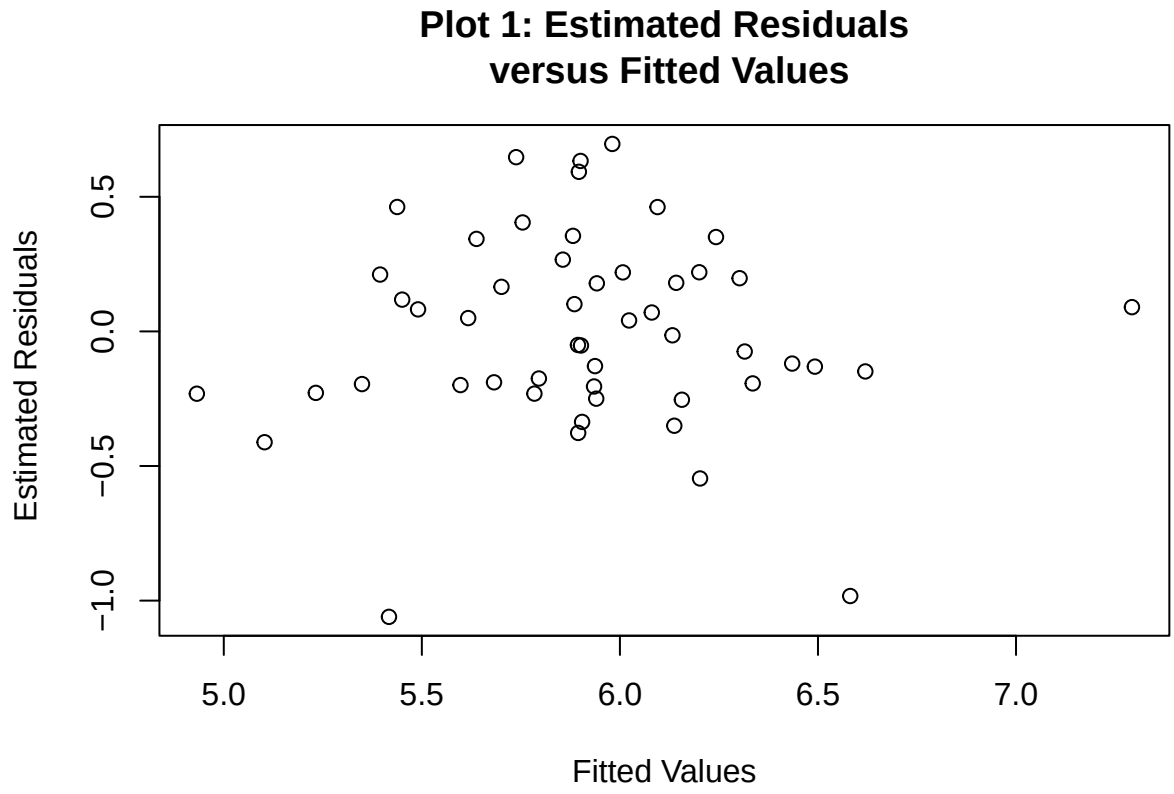
Conclusion of choice of model: From the analysis above, we should choose fit2 instead of fit1 because it better follows the assumptions made and therefore help us make accurate inferences of data.

```r
fit3=lm(log(violent)~metro+white+HS+poverty, data=UScrime05)
summary(fit3)
```

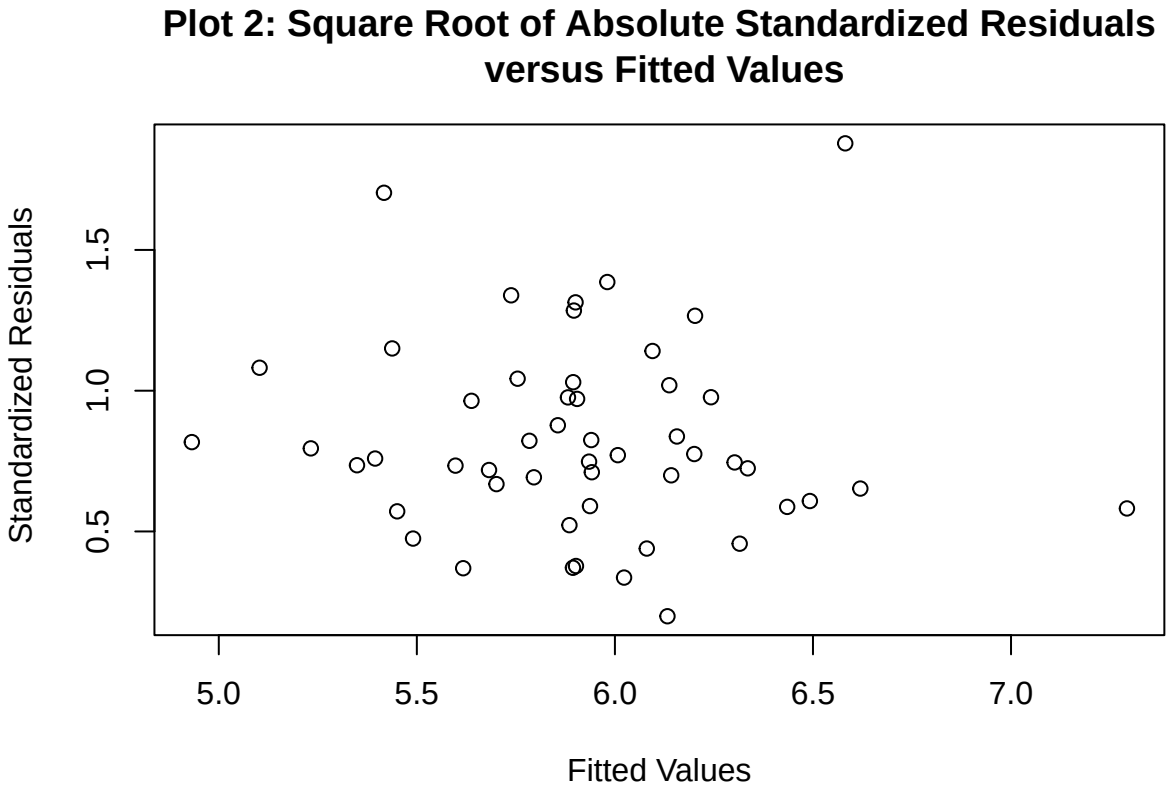
```
##
## Call:
## lm(formula = log(violent) ~ metro + white + HS + poverty, data = UScrime05)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06031 -0.20198 -0.01452  0.21519  0.69628
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.442532   2.092516   3.079   0.0035 **
## metro        0.017432   0.003878   4.495 4.68e-05 ***
## white       -0.008685   0.004557  -1.906   0.0629 .
## HS          -0.018688   0.020708  -0.902   0.3715
## poverty      0.056560   0.025498   2.218   0.0315 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3768 on 46 degrees of freedom
## Multiple R-squared:  0.5658, Adjusted R-squared:  0.528
## F-statistic: 14.98 on 4 and 46 DF,  p-value: 6.522e-08
```

From the summary above, both metro and poverty are having p-value less than 0.5, showing that they are significant factors of log(violent). Variable white and HS are having p-value to be larger than 0.05, showing that they are insignificant factors of log(violent). (Intercept) means the log(violent) is 6.442532 when metro, white, HS and poverty are all 0. The rates of change of log(violent) is 0.017432, -0.008685, -0.018688, 0.056560 under per unit change of metro, white, HS and poverty respectively.

```r
stres3 = sqrt(abs(rstandard(fit3)))
muhats3 = fitted(fit3)
plot(muhats3, fit3$residuals, xlab="Fitted Values",
     ylab=" Estimated Residuals",
     main="Plot 1: Estimated Residuals\n versus Fitted Values")
```

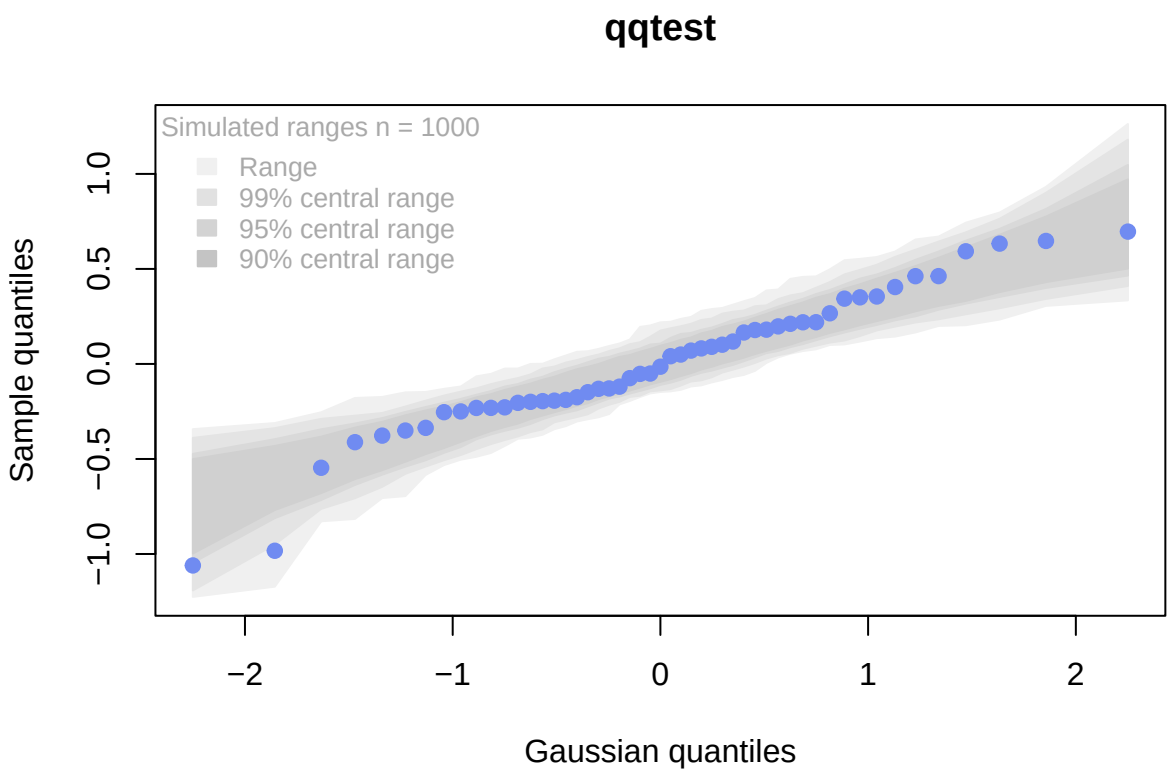



```r
plot(muhats3, stres3, xlab="Fitted Values",
     ylab=" Standardized Residuals",
     main="Plot 2: Square Root of Absolute Standardized Residuals\n versus Fitted Values")
```



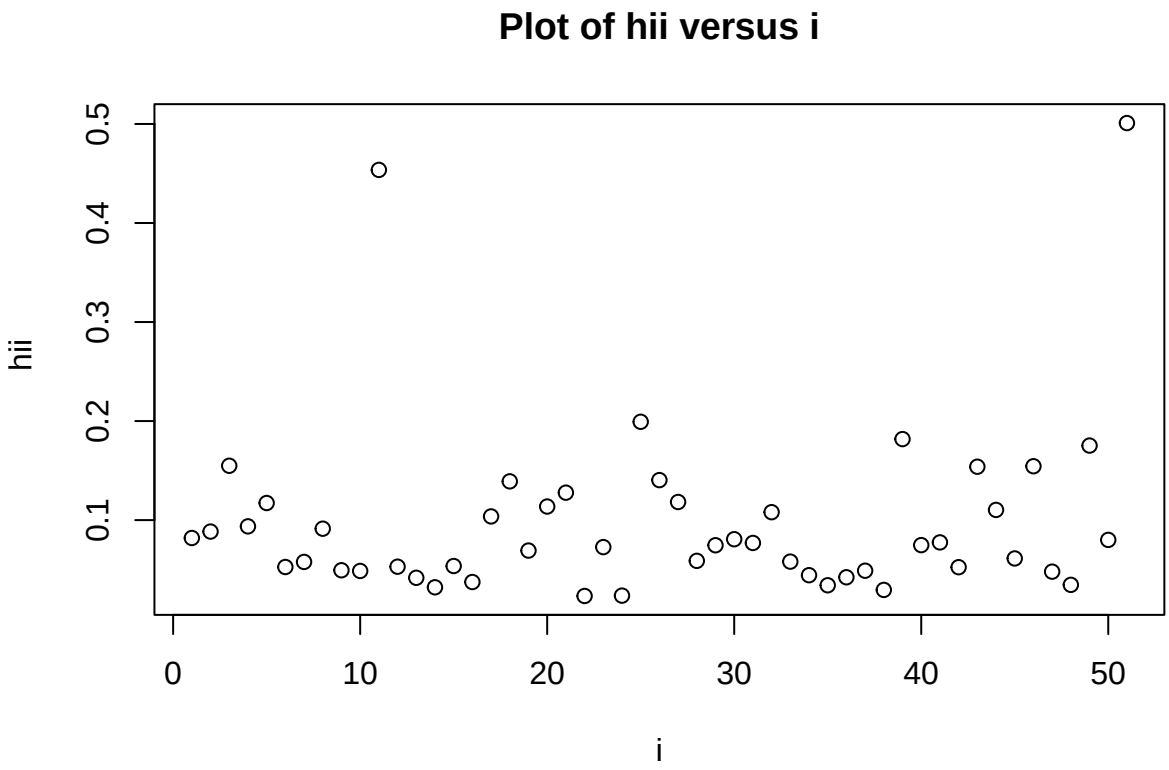

These two plots both showing residuals to be randomly distributed, and they are within 2 for standardised one. One special thing to mention is that there are two large residuals in plot2 that are larger than 1.5, but smaller than 2. So it still shows the heteroscedasticity of the data.

```
qqtest(fit3$residuals)
```



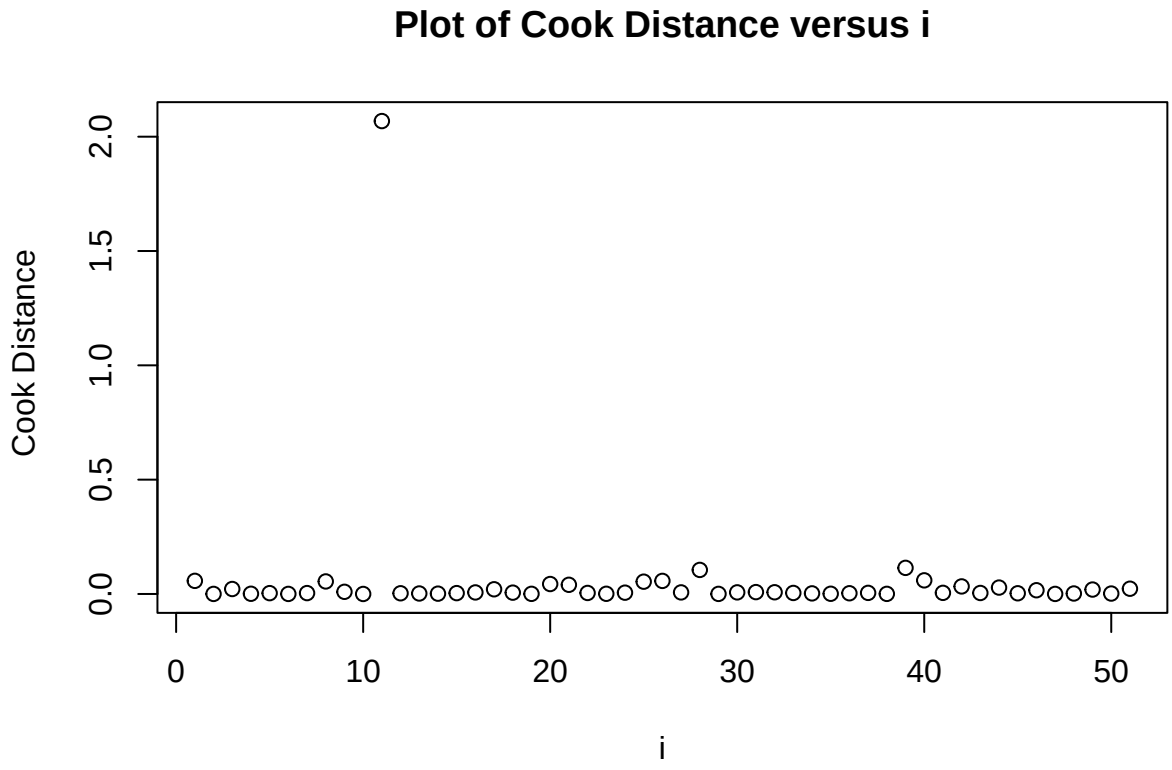

Since all residual points lie in shaded region, it shows that residuals are normally distributed.

```
hv2 = hatvalues(fit3)
plot(xseq, hv2, xlab="i",
     ylab=" hii",
     main="Plot of hii versus i")
```



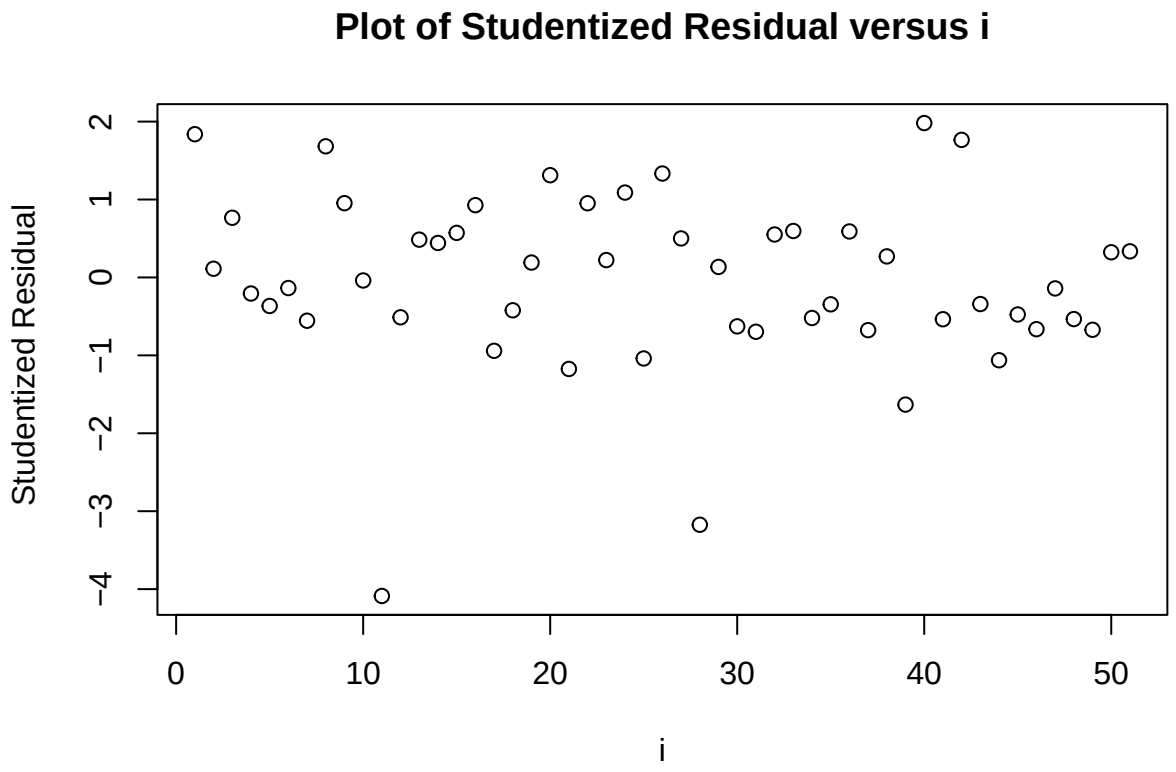

From the plot, it shows that District of Columbia(51) and Hawaii(11) have the potinetial for influence the fit.

```
cooksD =cooks.distance(fit3)
plot(xseq, cooksD, xlab="i",
     ylab=" Cook Distance",
     main="Plot of Cook Distance versus i")
```




From the plot, we can see that Hawaii had an actual influence on the fitted coefficient as its Cook Distance is greater than 2.

```r
rstu = rstudent(fit3)
plot(xseq, rstu, xlab="i",
     ylab=" Studentized Residual",
     main="Plot of Studentized Residual versus i")
```

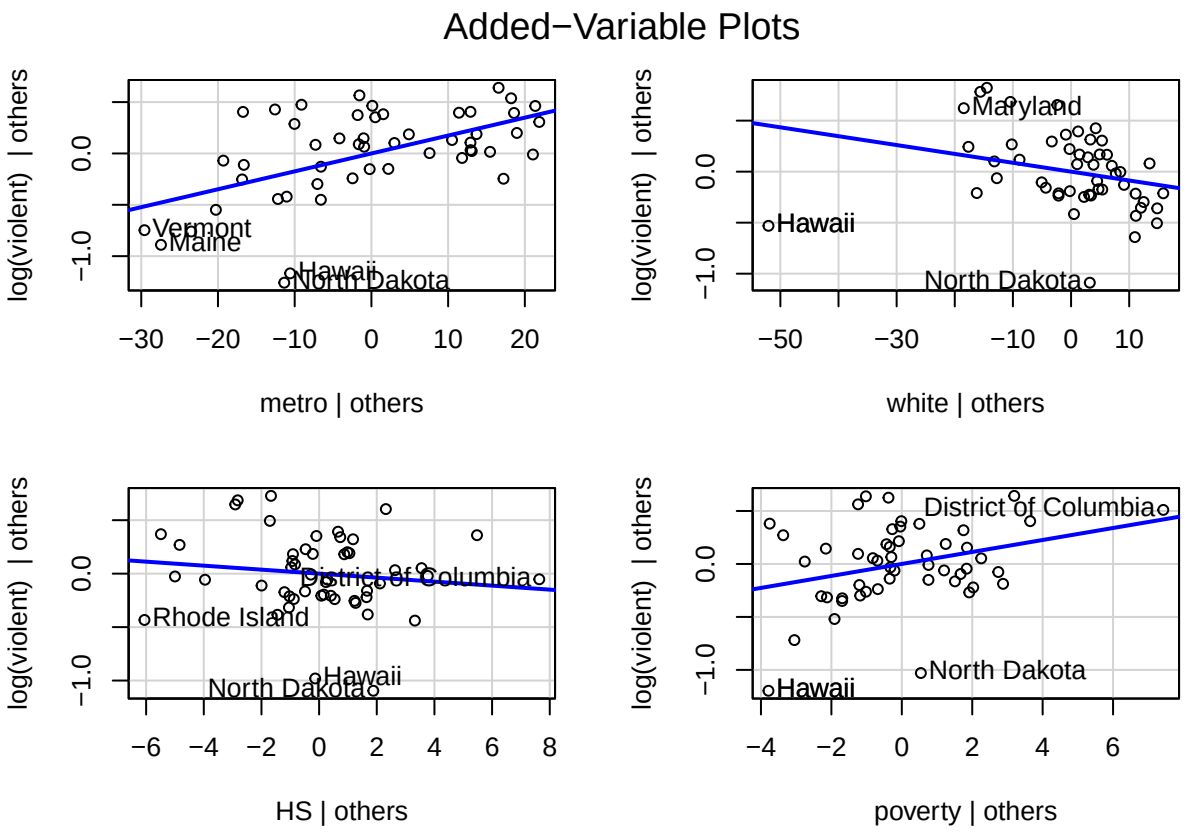



From the plot, we can see that Hawaii(11) and North Dakota(28) have studentized residual lying outside the range between -2 and 2.

```
library(car)
```

```
## Loading required package: carData
```

```
avPlots(fit3)
```




From the plots, we can see that most data follow the linear trend for metro and white, so these two variables should be considered. But data does not show any additional information for the prediction fo log(violent) for HS and poverty since data does not follow the linear trend in general, which is as expected. In addition, Hawaii and North Dakota are labelled in all of the plots, indicating that these two states are the influential points since they have influence to the variables we include in the model.