Crime statistics by state

```
load("UScrime05.rda")
```

Questions

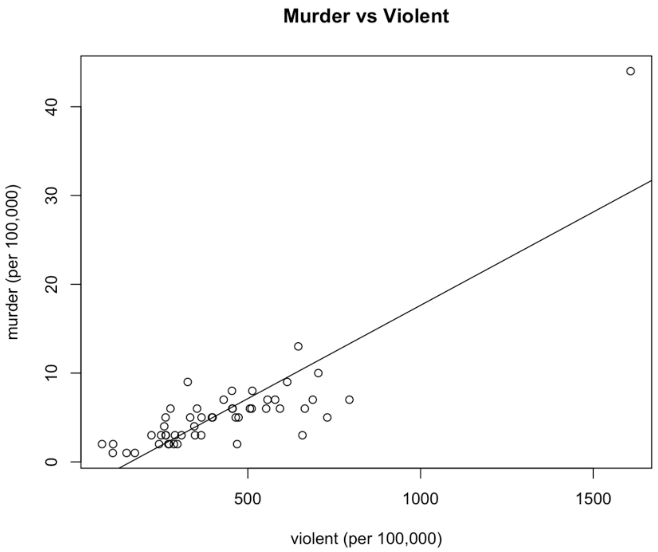
a. i. **Answer**

In the plot, see that as violence increase the murder rate also increases and there is an outlier with a very large violent crime rate and very large murder rate compared to the rest of the data. In the summmary, since the p value is small this means that the relationship between murder and violence is statistically significant and the R-squared value implies that 71.02% of the variance in murder rate is explained by violence.

```
# YOUR CODE HERE

fit1 = lm(UScrime05$murder~UScrime05$violent, data=UScrime05)

summary(fit1)
```

```
##
## Call:
## lm(formula = UScrime05$murder ~ UScrime05$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 ***
## UScrime05$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, main="Murder vs Violent", xlab="violent (per 100,000)", ylab="murder (per 100,000)")

abline(fit1)
```

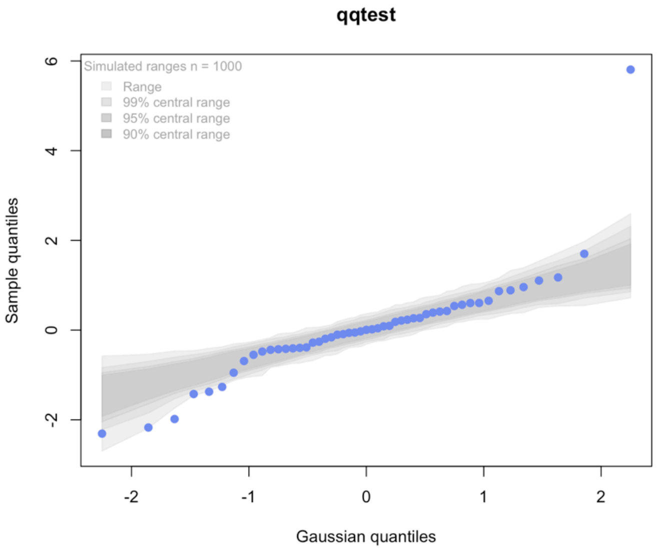


ii. **Answer**

This is the mathematical form of the residual estimates: $\hat{r}_i$ / $\sqrt{((sigmatild(i)^2) * (1 - hii))}$

From the plot, we see that the standardized residual estimates seem to follow a normal distribution in the, however, the points seem to deviate from the distribution as they head towards the tails. We start to see outliers at the tails with one data point very obviously being an outlier. So for the most part this indicates evidence against the hypothesis that the residuals are normally distributed.

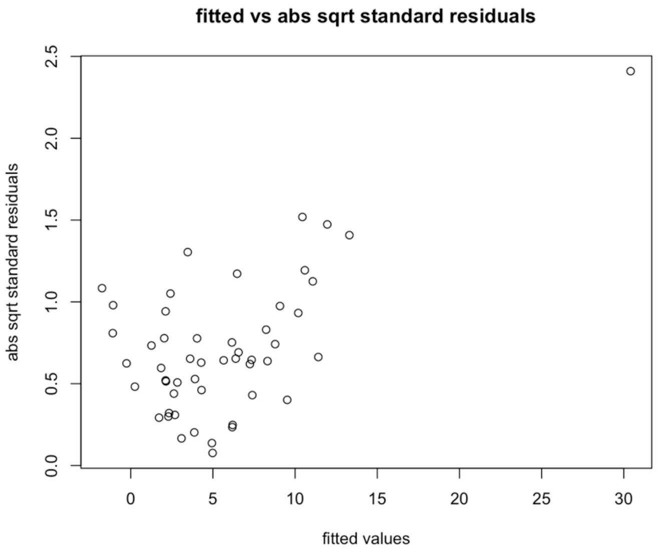
```
# YOUR CODE HERE
library("qqtest")

r_std = rstandard(fit1)
qqtest(r_std)
```



iii. **Answer**

From this plot, we can see that the spread of the residuals is not the same which means this shows evidence against the hypothesis of homoscedasticity. It is a good idea to plot this points to see the linearity and equal variance assumption.

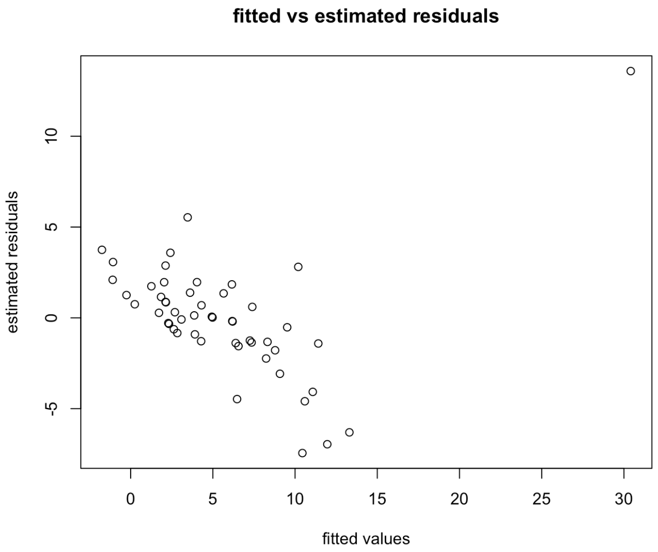
```
# YOUR CODE HERE

resid_abs_sqrt <- sqrt(abs(r_std))

plot(fitted(fit1), resid_abs_sqrt, main="fitted vs abs sqrt standard residuals",
xlab="fitted values", ylab="abs sqrt standard residuals")
```



iv. **Answer**

The simple estimated residuals are used because they help us to see whether the model accurately represents the variability in the data. This is because the value represents the difference between the observed and predicted values. From this plot, we see that as the fitted values increase the estimated residuals decrease for the most part. However, there is a an outlier that has a very high fitted value and a very high estimated residual compared to the rest of the data.

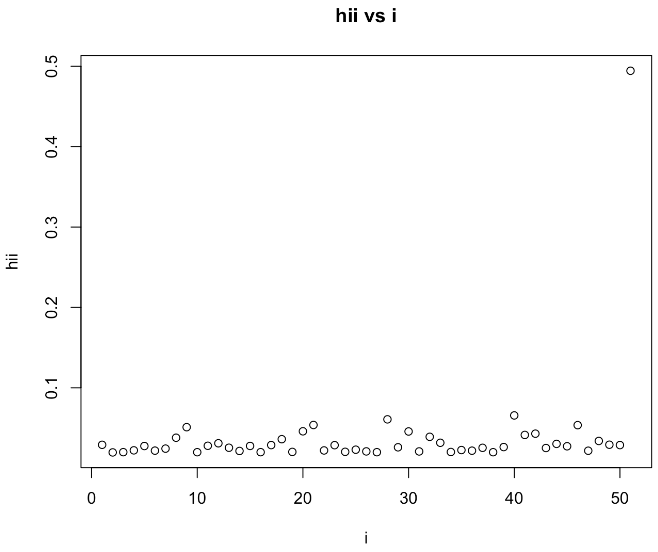
```
# YOUR CODE HERE

plot(fitted(fit1), resid(fit1), main="fitted vs estimated residuals", xlab="fitt
ed values", ylab="estimated residuals")
```



v. **Answer**

Points that have large hat values have a higher potential to affect the model. This is because they push more influence on the estimated coefficients. Specifically, District of Columbia has the highest potential to affect the model because it is an outlier. The hii value for it is around 0.5 which is much larger compared to the other hii values.

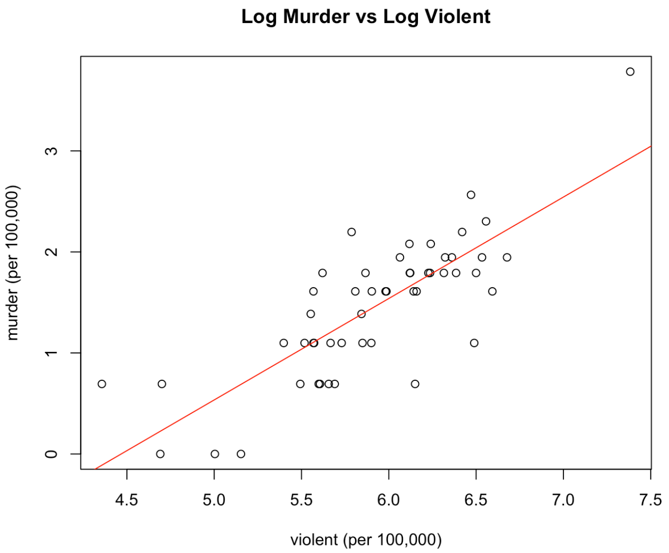
```
# YOUR CODE HERE

plot(hatvalues(fit1), main="hii vs i",
   xlab = "i", ylab = "hii")
```



b. i. **Answer**

As the violent crime rate log increases, the murder rate log also increases. There is a higher concentration of data points in the middle and a few data points on the left tail. There is an outlier which has a very high violent crime rate log and a very high murder rate log compared to the other data points.

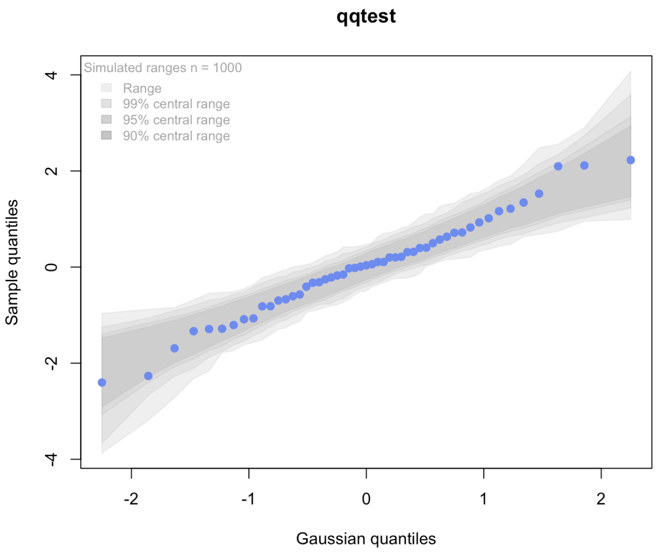
```
# YOUR CODE HERE

fit2 = lm(log(UScrime05$murder)~log(UScrime05$violent), data=UScrime05)

plot(log(UScrime05$violent), log(UScrime05$murder), main="Log Murder vs Log Viol
ent", xlab="violent (per 100,000)", ylab="murder (per 100,000)")
abline(fit2, col="red")
```



ii. **Answer**

From the plot, the standardized residual estimates seem to follow a normal distribution in the middle. At the tails, the data points tend to breach the 90% central range while still staing within the 99% central range. This implies that there is evidence for the hypothesis that the residuals are normally distributed.

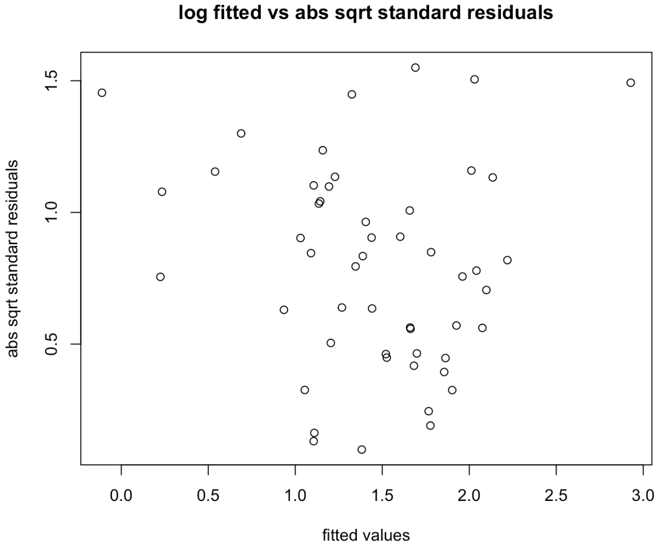
```
# YOUR CODE HERE

r_std2 <- rstandard(fit2)
qqtest(r_std2)
```



iii. **Answer**

From the plot, we can see that there is more evidence for the hypothesis of homoscedasticity compared to fit1. This is because here the spread of all the data points are relatively similar and the spread of all the data points in fit1 were different.

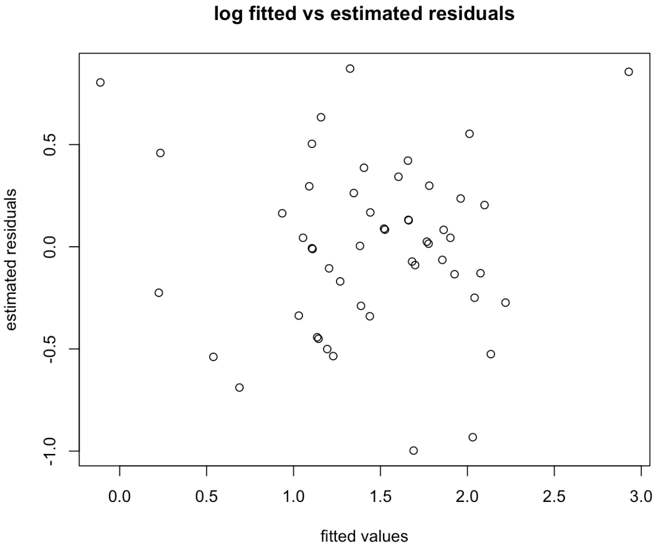
```
# YOUR CODE HERE

r_abs_sqrt2 <- sqrt(abs(r_std2))

plot(fitted(fit2), r_abs_sqrt2, main="log fitted vs abs sqrt standard residuals", xlab="fitted values", ylab="abs sqrt standard residuals")
```



iv. **Answer**

A possible structure that may be missing is a horizontal line. This would suggest that the model is accurate.

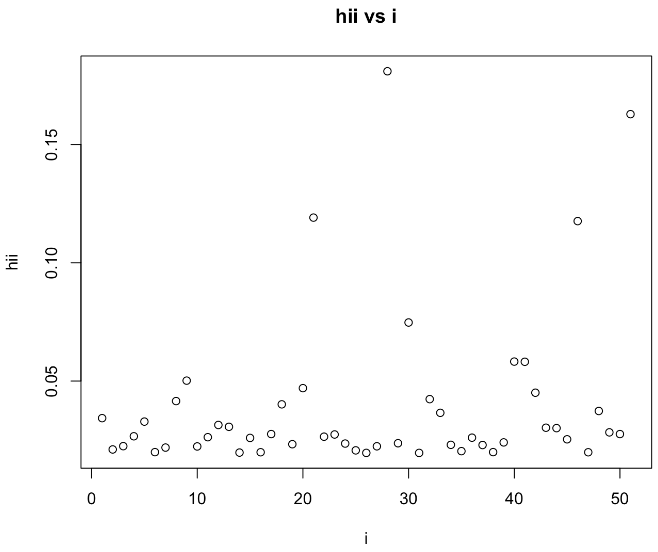
```
# YOUR CODE HERE

plot(fitted(fit2), resid(fit2), main="log fitted vs estimated residuals",xlab="f
itted values", ylab="estimated residuals")
```



v. **Answer**

The potential influence of the District of Columbia did change. From this plot, we see that instead North Dakota has a stronger potential influence now. This has occured because the log transformation compressed the range of the data and decreased the impact of the outliers that we saw before.

```
# YOUR CODE HERE

plot(hatvalues(fit2), main="hii vs i", xlab = "i", ylab = "hii")
```



vi. **Answer**

These values tell us that 71% of the variation in murder rate can be explained by fit1 and 64% of the variation in murder rate can be explained by fit2. This fit1 seems to be better simply by the rsquared value, but it is not enough to make a decision.

I would recommend that fit2 is a better model because although both models appear to follow a linear pattern in the non-residual scatter plots, we see that the spread of the data points in fit2 are similar in the residual scatter plot. From the qqtest, the some of the data points for fit1 are outside of the intervals wherase in fit2, at least all points are within the range.

```
# YOUR CODE HERE

rsquared1 <- summary(fit1)$r.squared
rsquared2 <- summary(fit2)$r.squared
```

c. i. **Answer**

From the summary, we can see that the p value for the metro variable is very small which implies that there is no evidence to support the hypothesis that the violent crime rate log is higher in metro areas compared to in non-metro areas. We can see that the p value for the white variable is a little small which implies that there is little evidence to support the hypothesis that the violent crime rate log is less in the population of an area that is white. We can see that the p value for the HS variable is a little larger which implies that there is some evidence to support the hypothesis that the violent crime rate log is less in the population of an area who have graduated high school. We can see that the p value for the poverty variable is small which suggests little to no evidence to support the hypothesis that the violent crime rate log is higher in the population of an area that has is below the poverty level.

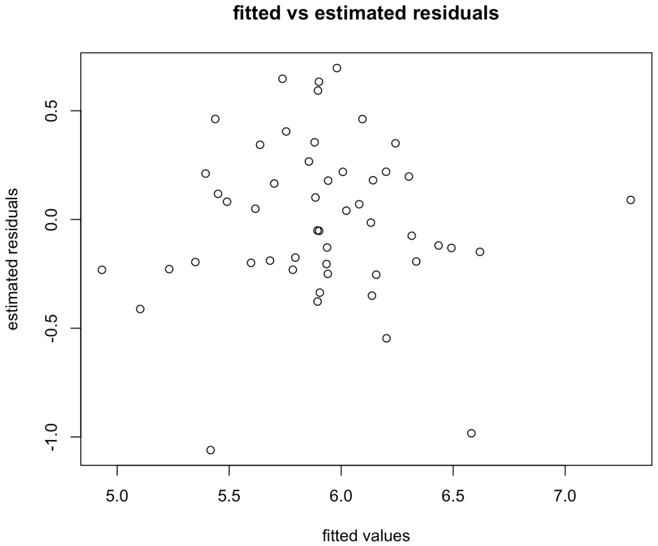
```
# YOUR CODE HERE

fit3 <- lm(log(UScrime05$violent) ~ UScrime05$metro + UScrime05$white + UScrime0
5$HS + UScrime05$poverty)
summary(fit3)
```

```
##
## Call:
## lm(formula = log(UScrime05$violent) ~ UScrime05$metro + UScrime05$white +
##     UScrime05$HS + UScrime05$poverty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06031 -0.20198 -0.01452  0.21519  0.69628
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.442532   2.092516   3.079   0.0035 **
## UScrime05$metro    0.017432   0.003878   4.495 4.68e-05 ***
## UScrime05$white   -0.008685   0.004557  -1.906   0.0629 .
## UScrime05$HS      -0.018688   0.020708  -0.902   0.3715
## UScrime05$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
```

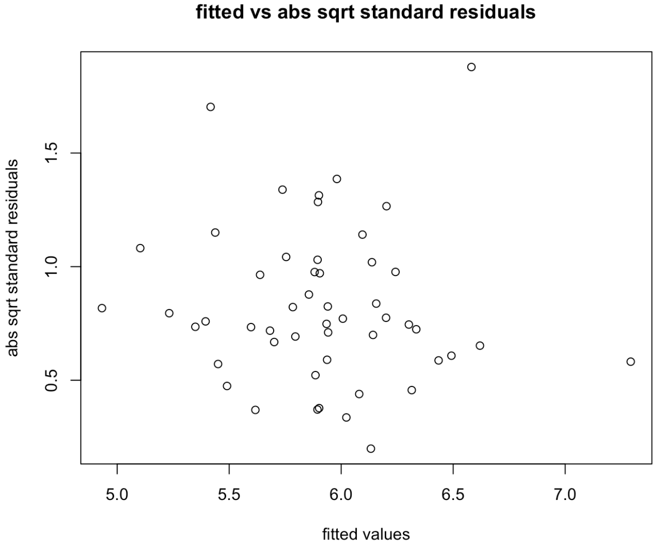
ii. **Answer**

In plot 1, we see that the residuals have an average that is closer to 0 more so than in plot 2. We also see that there are two large residuals in plot 1. In plot 2, we see the spread is even across the data points more so than in plot 1.

```
# YOUR CODE HERE

plot(fitted(fit3), resid(fit3), main="fitted vs estimated residuals", xlab="fitt
ed values", ylab="estimated residuals")
```



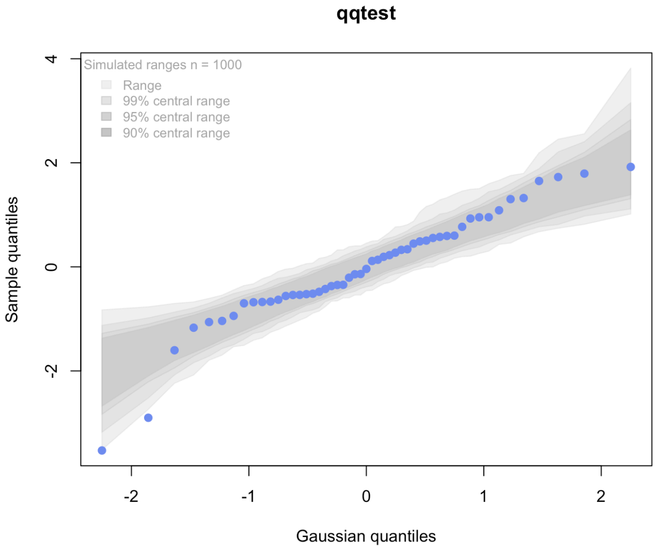
```
r_std3 <- rstandard(fit3)
r_abs_sqrt3 <- sqrt(abs(r_std3))

plot(fitted(fit3), r_abs_sqrt3, main="fitted vs abs sqrt standard residuals", xlab="fitted values", ylab="abs sqrt standard residuals")
```



iii. **Answer**

From the plot, we see that the points in the middle seem to follow a normal distribution. The points at the tails seem to deviate from the normal distribution as they start to breach the ranges. There is even one point that is completely outside of the range.

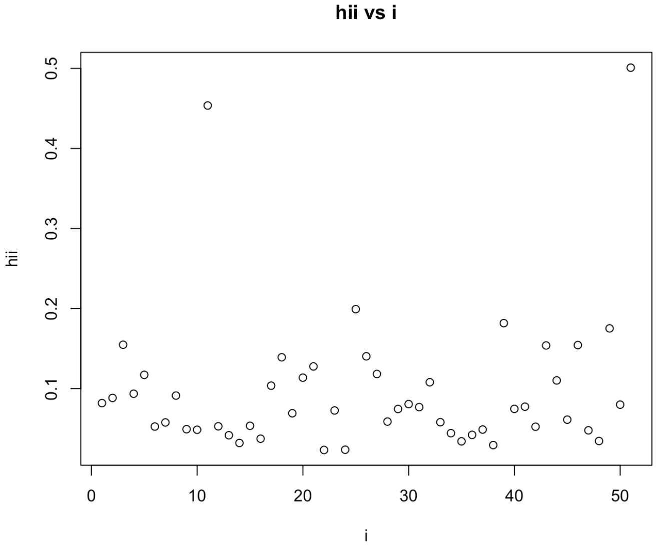
```
# YOUR CODE HERE

qqtest(r_std3)
```



iv. **Answer**

The data points that seem to have a high potential influence are District of Columbia and Hawaii.

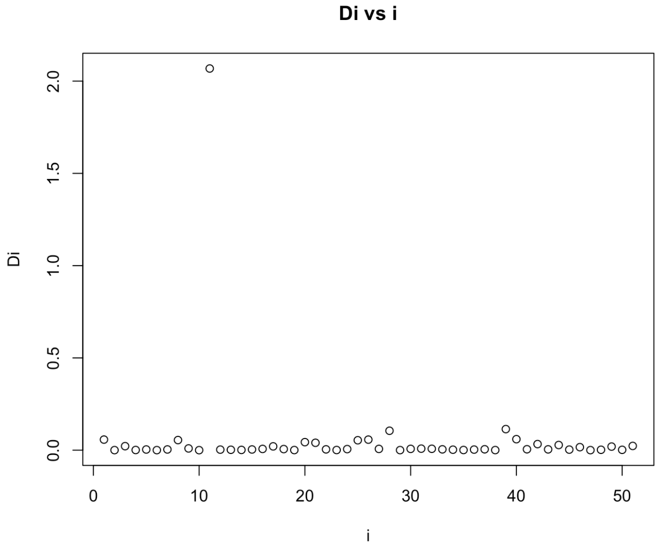
```
# YOUR CODE HERE

plot(hatvalues(fit3), main="hii vs i", xlab = "i", ylab = "hii")
```



v. **Answer**

From the plot, we see that Hawaii had an influence on the fit. This is because its corresponding Cook's distance value is very far from the rest of the points.

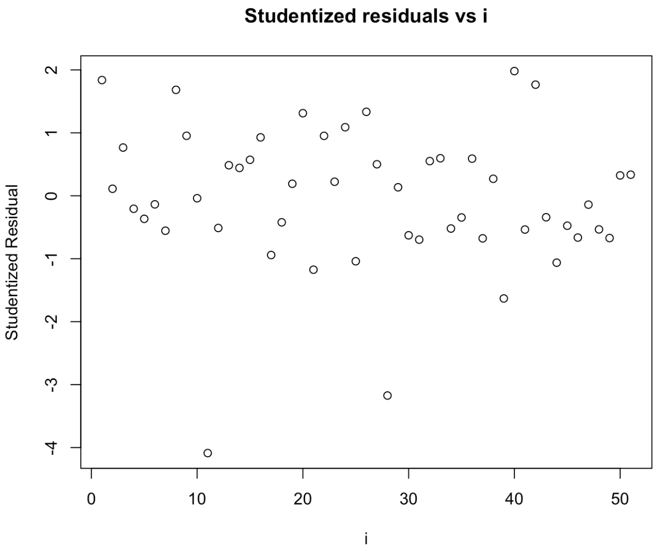
```
# YOUR CODE HERE

plot(cooks.distance(fit3), main="Di vs i", xlab = "i", ylab = "Di")
```



vi. **Answer**

From this plot, we see that Hawaii and North Dakota are outliers. This is because their studentized residual values are farther from the rest which seem to spread evenly and within a certain range.

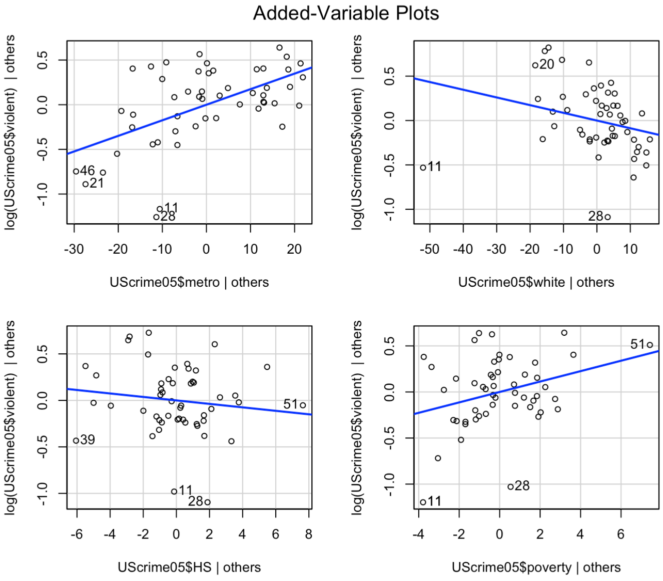
```
# YOUR CODE HERE

plot(rstudent(fit3), main="Studentized residuals vs i", xlab = "i", ylab = "Studentized Residual")
```



vii. **Answer**

From the plots, we see that as the metro values increase so does the violent crime rate log. We see that as the white values increase, the violent crime rate log decreases. We see that as the highschool values increase, the violent crime rate log decreases. We can see that as the poverty values increase so does the violent crime rate log. The line in poverty and metro seem to have a higher slope than the other two. The data points in metro seem to the most linear shape whereas the data points in white do not. We also see that the outliers in the white plot are much stronger compared to the others.

```
# YOUR CODE HERE

library("car")
avPlots(fit3)
```


Added-Variable Plots
log(UScrime05$violent) | others
0.5
0.0
-0.5
-1.0
46
21
11
28
-30
-20
-10
0
10
20
UScrime05$metro | others
log(UScrime05$violent) | others
0.5
0.0
-0.5
-1.0
20
11
28
-50
-40
-30
-20
-10
0
10
UScrime05$white | others
log(UScrime05$violent) | others
0.5
0.0
-0.5
-1.0
39
11
28
51
-6
-4
-2
0
2
4
6
8
UScrime05$HS | others
log(UScrime05$violent) | others
0.5
0.0
-0.5
-1.0
51
28
11
-4
-2
0
2
4
6
UScrime05$poverty | others