STAT 210
Applied Statistics and Data Analysis
Linear Regression VI:
Influential Points and Transformations

Joaquin Ortega

Influential and Atypical Points

Influential and Atypical Points

Atypical points are data that have large residuals. We classify an observation as atypical if its residual is large in relation to the residuals of the rest of the observations.

Influential points, on the other hand, are points that have a strong influence on the model. By this, we mean that if the model is fitted excluding these points, the model changes substantially.

Let's see an example of the effect of influential points on regression, taken from Chatterjee, Hadi, & Price's book[1].

The success of a television program is partially evaluated by its rating, which is an attempt to measure the program's ability to attract and to retain an audience with an index that varies between 1 and 10.

---

[1] Chatterjee, Hadi & Price *Regression Analysis by Example, 3rd Edition*, Wiley, 1999

Influential and Atypical Points

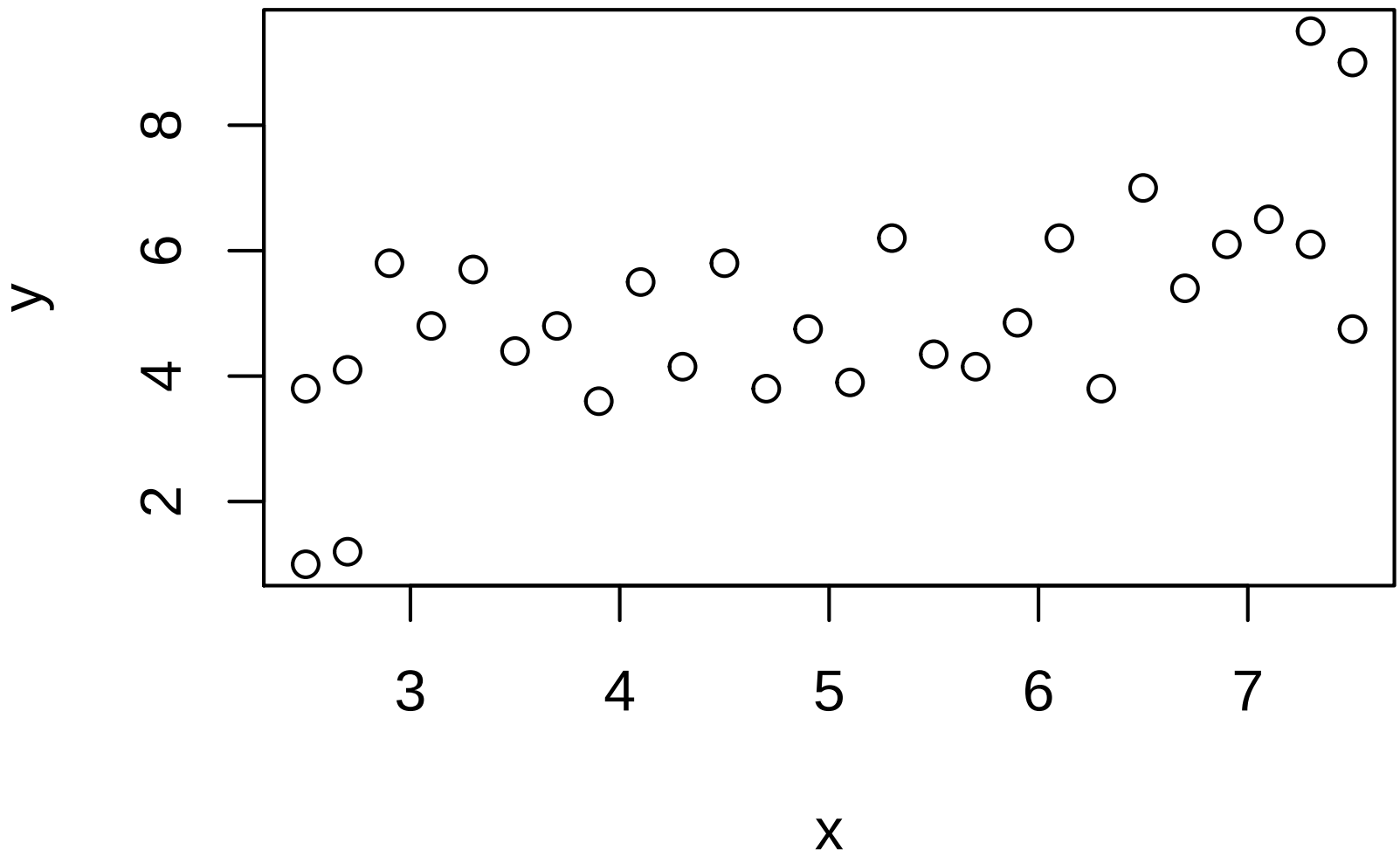
The following data supposedly comes from a study to measure the effect that the rating of the previous program has on the audience of a newscast, seeking to explore the idea that part of the audience of the first program 'stays' for the newscast.

To quantify this effect, a sample of ratings was taken from different localities and times.

The observations are in the `TV` file and consist of the variable y that represents the rating of the newscast and variable x that represents the rating of the previous program to the news. We start by graphing the data

Influential and Atypical Points

```
tv <- read.csv('TV.csv')
plot(tv)
```



The graph shows a possible linear relationship between the variables with a positive slope.

Influential and Atypical Points

We estimate the parameters for the regression.

```
fittv <- lm(tv$y ~ tv$x)
summary(fittv)
```

```
##
## Call:
## lm(formula = tv$y ~ tv$x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36994 -0.95755 -0.06405  0.96824  2.93634
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7065     0.8172   2.088 0.045977 *
## tv$x          0.6654     0.1552   4.287 0.000194 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 28 degrees of freedom
## Multiple R-squared:  0.3963, Adjusted R-squared:  0.3747
## F-statistic: 18.38 on 1 and 28 DF,  p-value: 0.0001939
```

Influential and Atypical Points

The results indicate that the previous show's rating can explain nearly 40% of the variation in news ratings.

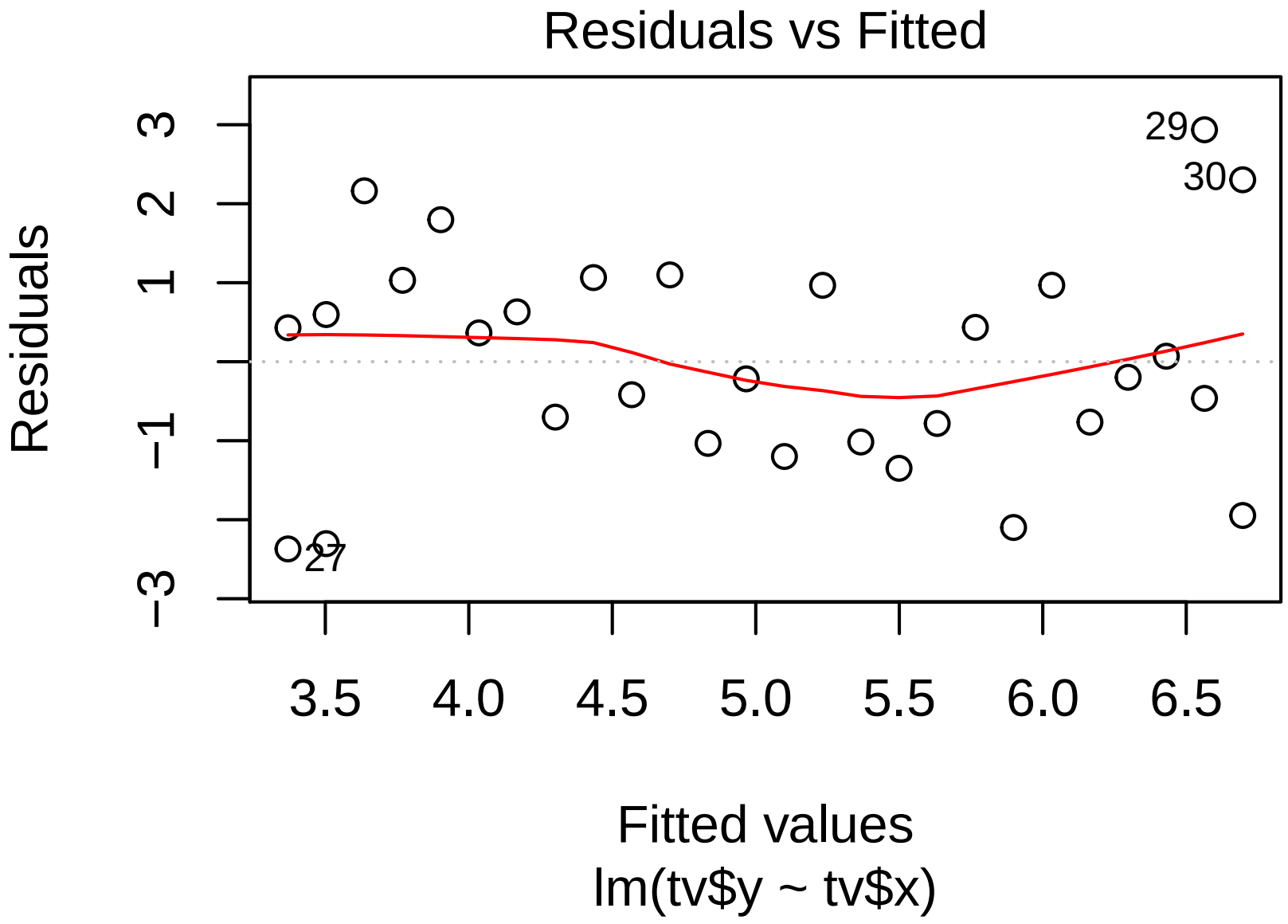
Moreover, for every point that the previous program's rating increases, we expect the newscast's rating to increase by 0.66.

These assessments seem relevant, but are only valid as long as there are no major violations of the model's hypotheses.

To verify if this is the situation, we analyze the residuals of the model through a series of graphs.

Influential and Atypical Points

```
plot(fittv, which=1)
```



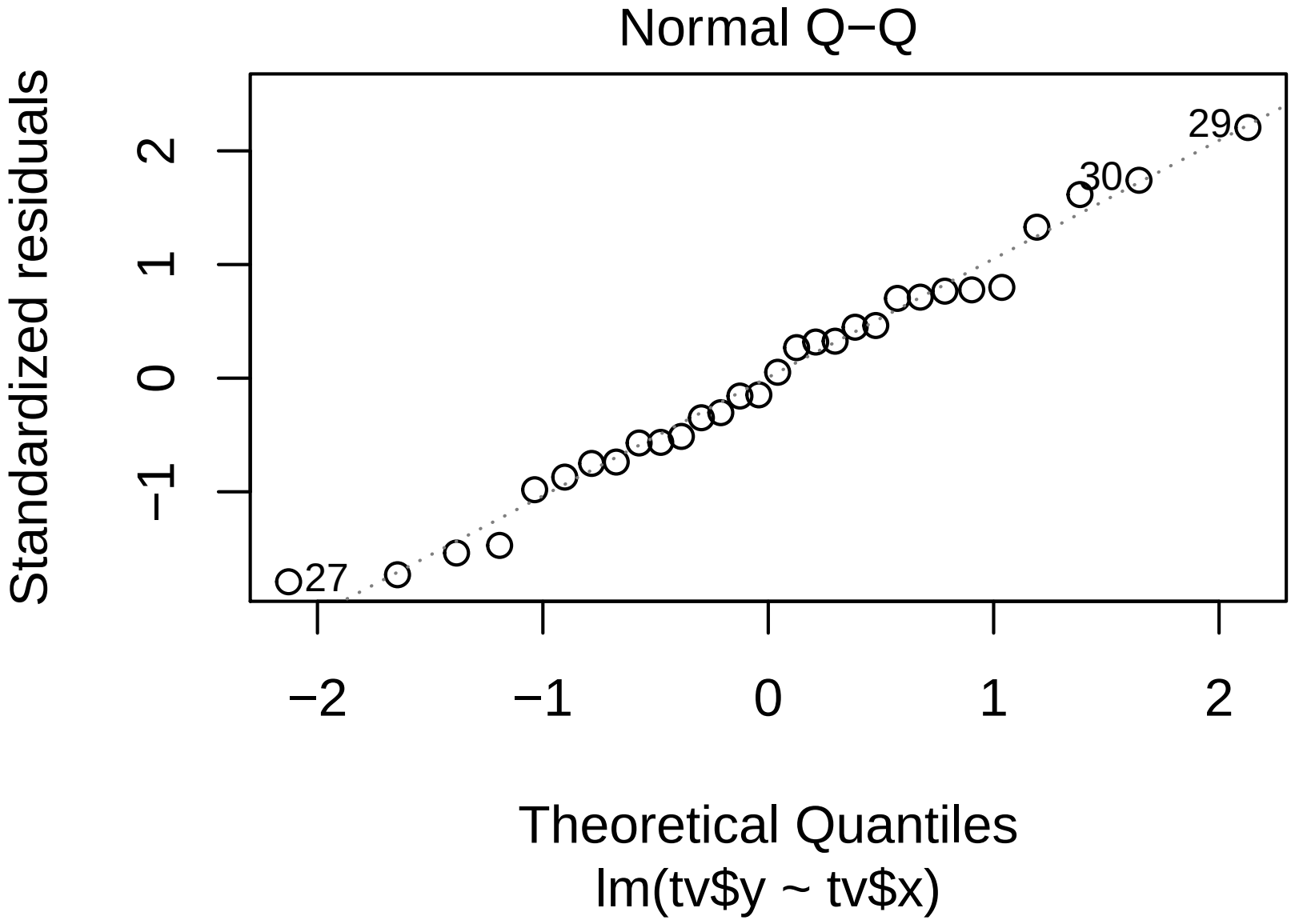
Influential and Atypical Points

In this plot, we see that there are four points, two on the upper right-hand corner with numbers 29 and 30, and two on the lower left-hand corner, one of which is numbered 27, which seem to 'tilt' the graph because the rest of the points seem to follow a decreasing pattern with a small negative slope.

Next, we plot the other three diagnostic graphs we have considered before.

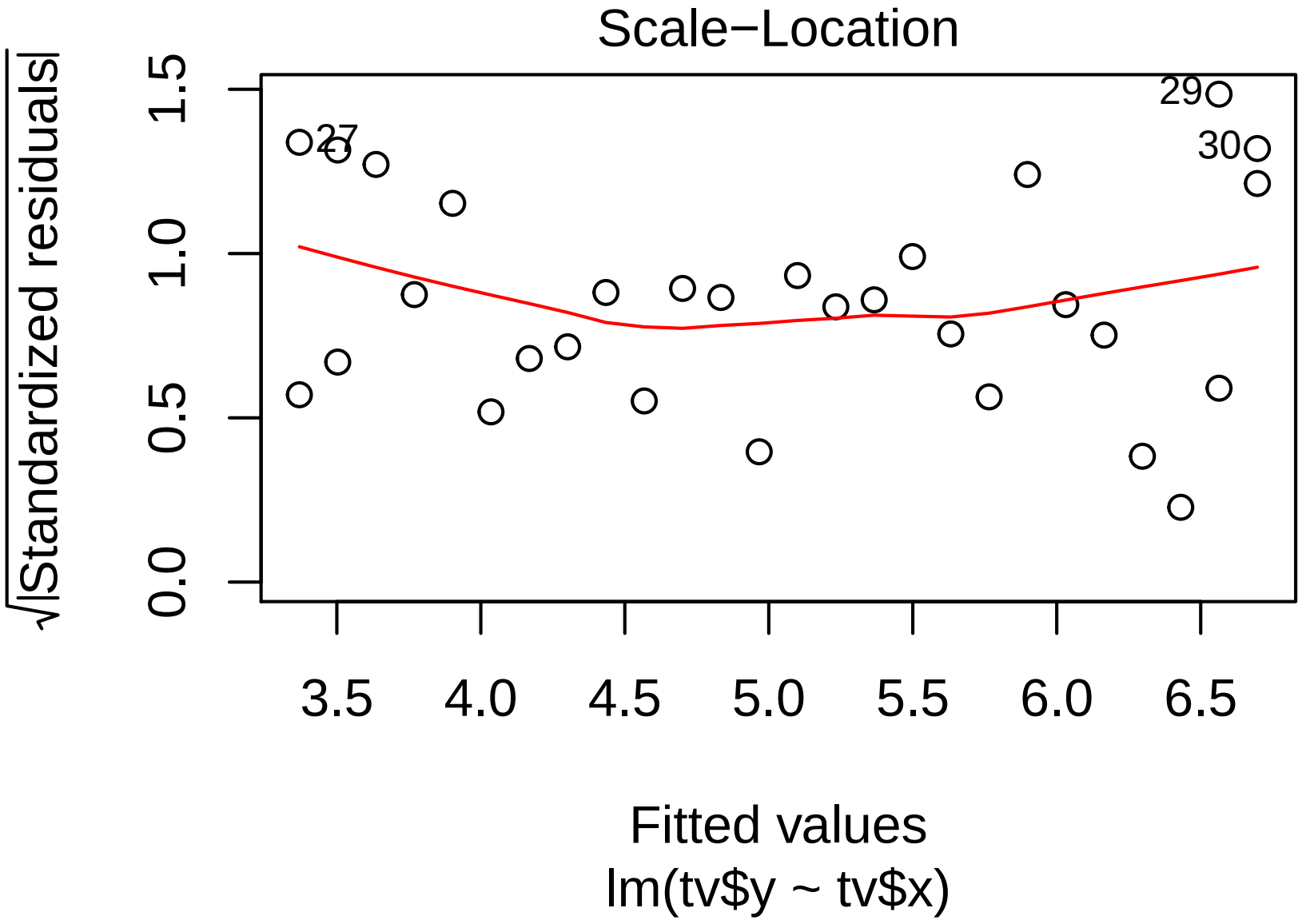
Influential and Atypical Points

```
plot(fittv, which=2)
```



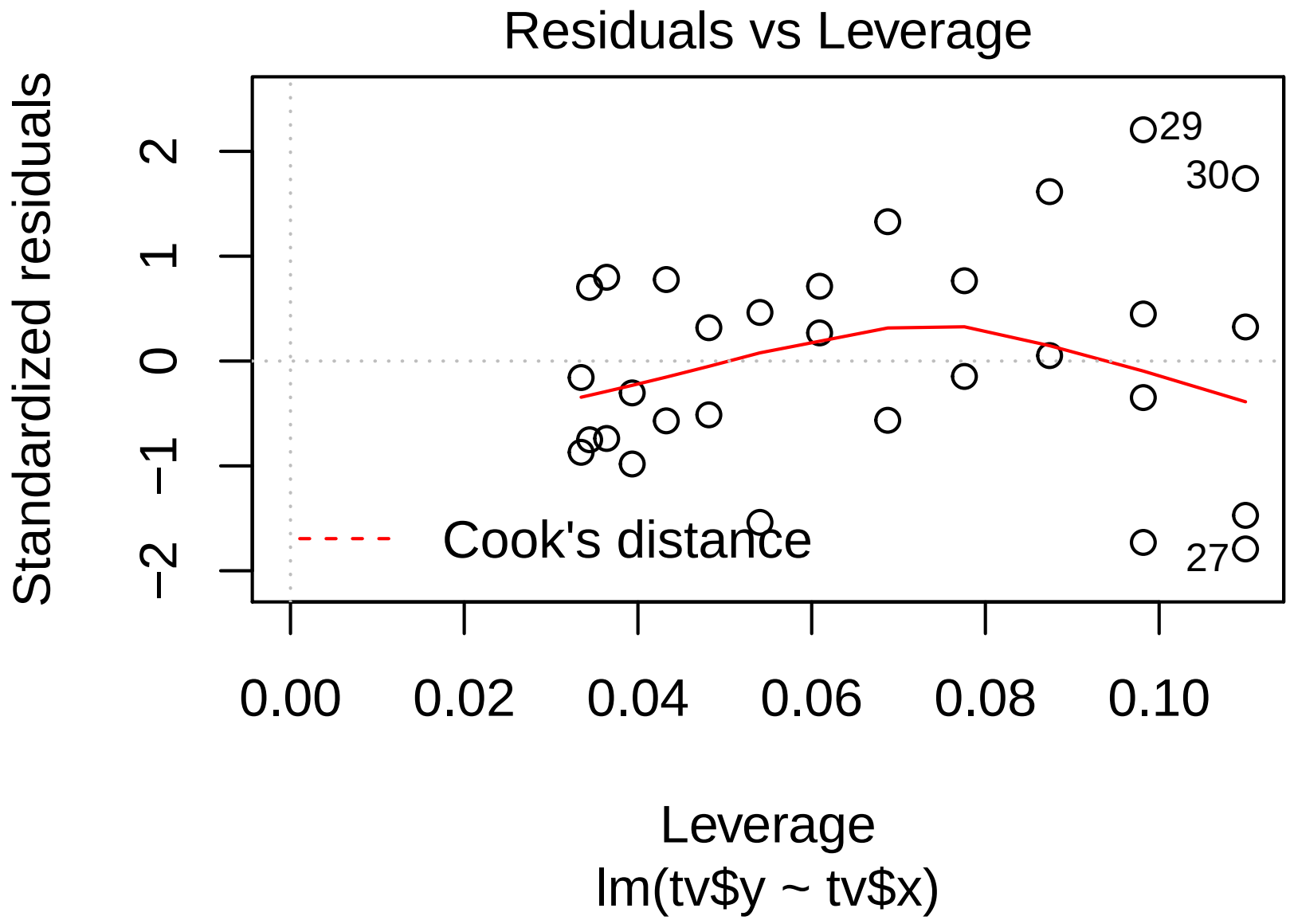
Influential and Atypical Points

```
plot(fittv, which=3)
```



Influential and Atypical Points

```
plot(fittv, which=5)
```



Influential and Atypical Points

The fit in the quantile plot is very good, and in general, the graphs are adequate, but the points are always flagged as potentially influential points.

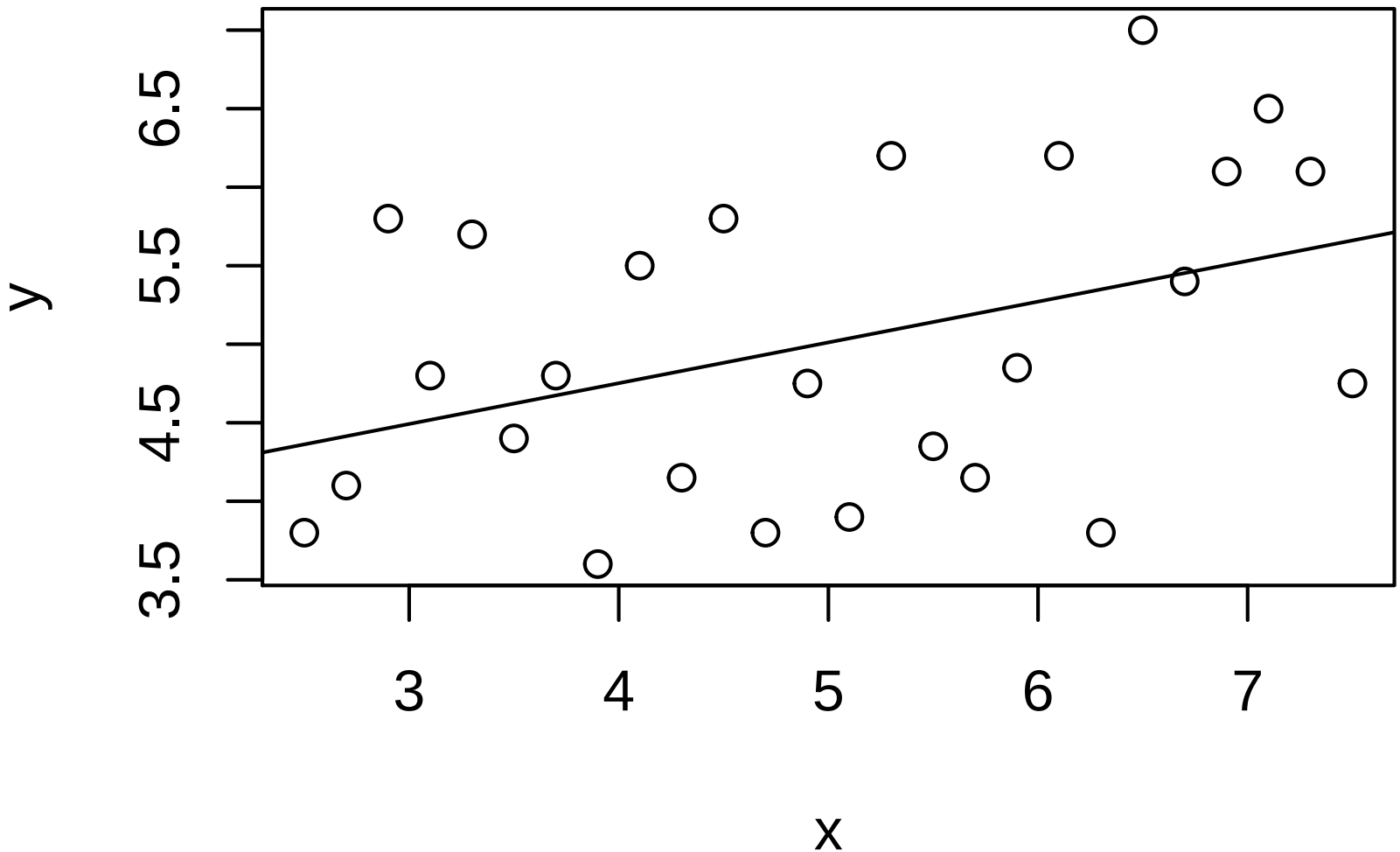
Going back to the initial graph, we see that these points also appear on the lower left-hand and upper right-hand corners, and we observe that if they are excluded, the increasing linear relation seems less obvious.

These points seem to have a significant influence on the slope. They should be carefully investigated to verify whether they are mistakes or, on the contrary, genuine data points that could have interesting information regarding this problem.

Influential and Atypical Points

As an exercise we fit the model excluding these four points.

```
tv2 <- tv[1:26,]; plot(tv2)
fittv2 <- lm(tv2$y ~ tv2$x)
abline(fittv2)
```



Influential and Atypical Points

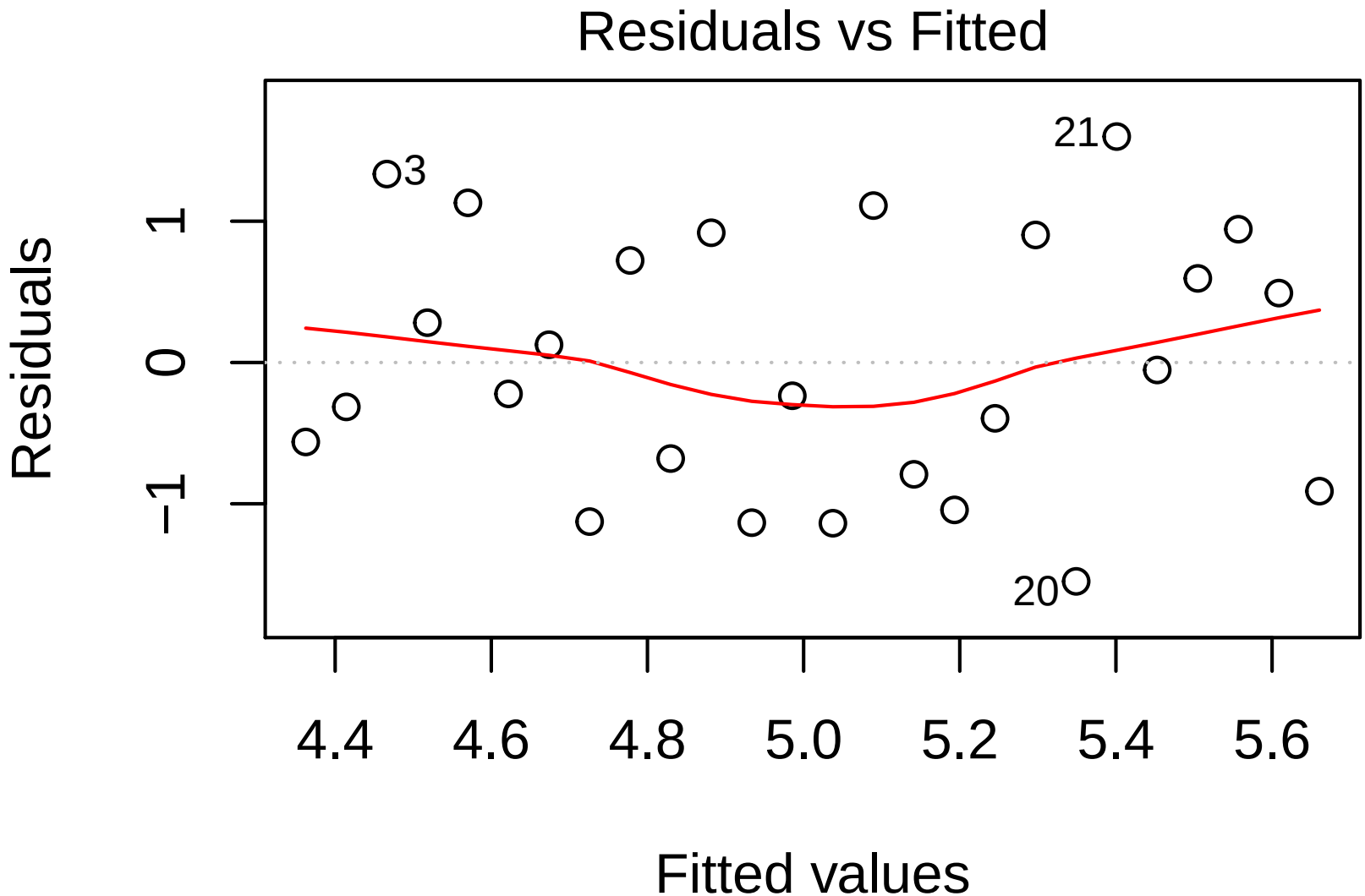
```
summary(fittv2)
```

```
##
## Call:
## lm(formula = tv2$y ~ tv2$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5491 -0.7635 -0.1375  0.8577  1.5990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7132     0.6314   5.881 4.56e-06 ***
## tv2$x         0.2597     0.1209   2.147   0.0421 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9251 on 24 degrees of freedom
## Multiple R-squared:  0.1611, Adjusted R-squared:  0.1262
## F-statistic: 4.609 on 1 and 24 DF,  p-value: 0.04211
```

Influential and Atypical Points

The results change radically. The slope, which is reduced to 0.26, is moderately significant, and the R^2 goes down to 0.1611.

```
plot(fittv2)
```



Transformed Data

Transformed Data

The data in the `Bacteria` file represent the number (in hundreds) of marine bacteria that survived 200 kilovolt X-ray exposure for periods ranging from 1 to 15 6-minute intervals.

The experiment was conducted to test the hypothesis that bacterial deaths occur when their 'vital center' is struck by a ray. This type of bacteria does not form groups or chains, so the number of bacteria can be estimated by plate counts.

If the theory is correct, the logarithm of the number of survivors must have a linear relationship to the length of the exposure.

Transformed Data

If n_t represents the number of surviving bacteria at time t

$$n_t = n_0 e^{\beta t}, \quad t > 0,$$

where n_0 and β are the model parameters and have simple interpretations: n_0 is the number of bacteria at the beginning of the experiment, and β is the rate of destruction or death of the bacteria. Taking logarithms

$$\log n_t = \log n_0 + \beta t = \alpha + \beta t$$

where $\alpha = \log n_0$, and we see that this is a linear function of t. If we add a random error to the model, we get

$$\log n_t = \alpha + \beta t + \epsilon_t$$

and we can now fit a regression model.

Transformed Data

If we go back to the original (exponential) model the error appears multiplicatively:

$$n_t = n_0 e^{\beta t} u_t,$$

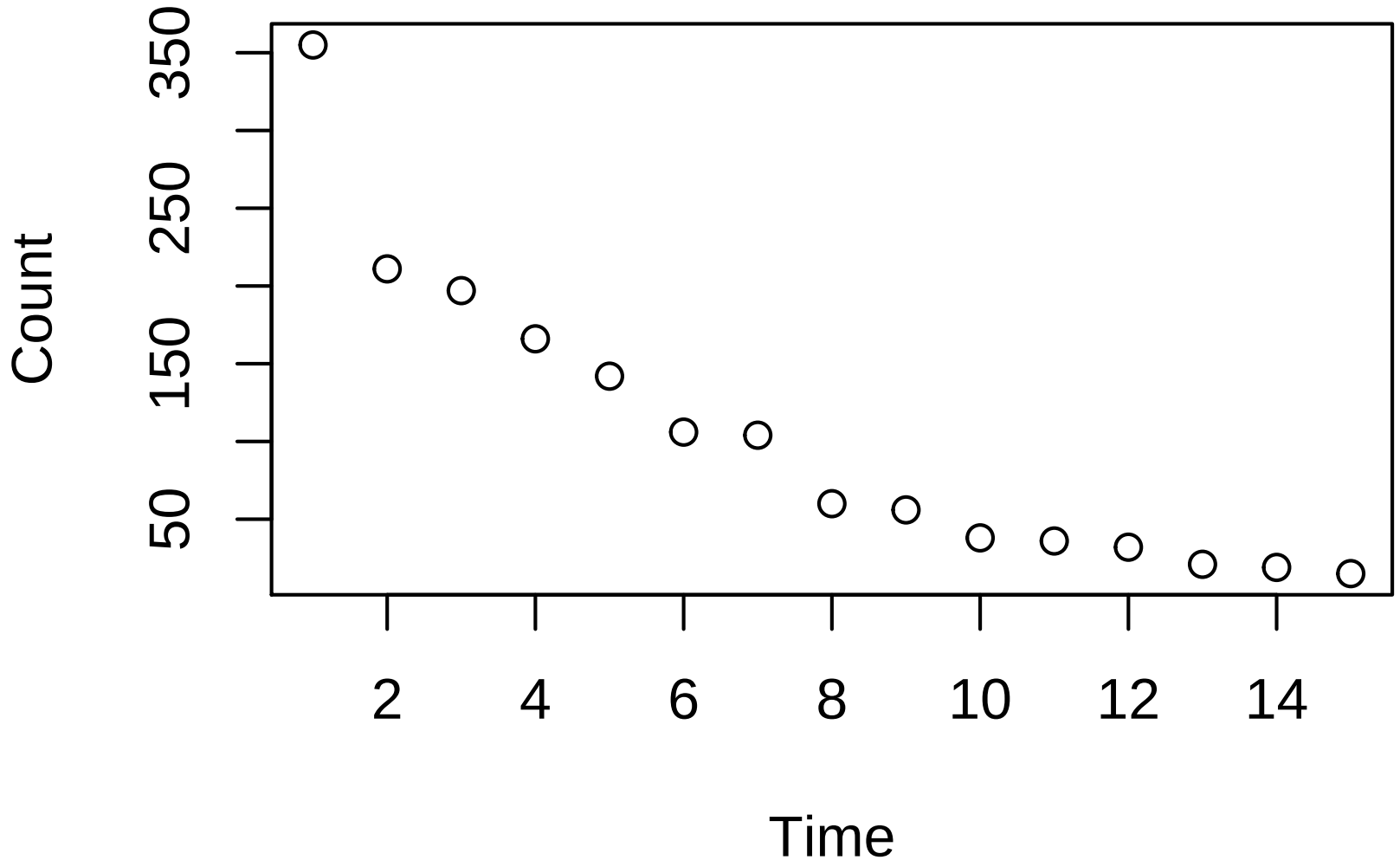
where $u_t = e^{\epsilon_t}$ is a multiplicative error.

The model assumes that $\epsilon_t = \log u_t$ has a normal distribution and therefore u_t must have lognormal distribution.

Transformed Data

We start graphing the data

```
Bacteria <- read.csv('Bacteria.csv')
plot(Bacteria)
```



Transformed Data

The graph suggests a non-linear relationship between the two variables.

However, we proceed to fit a linear model to study the consequences. The model is

$$n_t = \alpha + \beta t + \epsilon_t,$$

Transformed Data

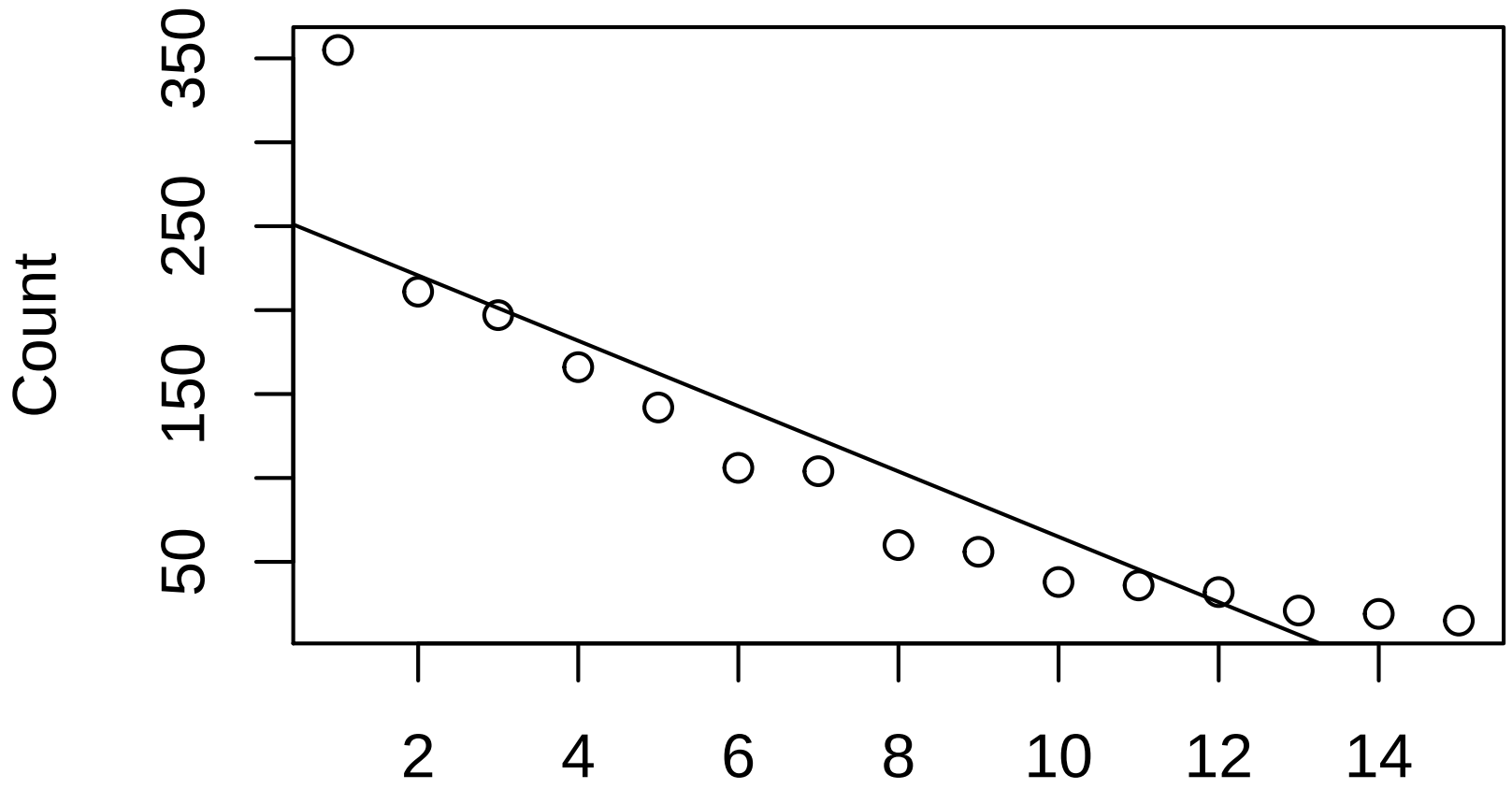
```r
attach(Bacteria)
fitbac <- lm(Count ~ Time)
summary(fitbac)
```

```
##
## Call:
## lm(formula = Count ~ Time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.867 -23.599  -9.652  10.223 114.883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   259.58      22.73  11.420 3.78e-08 ***
## Time          -19.46       2.50  -7.786 3.01e-06 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.83 on 13 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8098
## F-statistic: 60.62 on 1 and 13 DF,  p-value: 3.006e-06
```

Transformed Data

Although the regression coefficient is significant, and we have a high value for R^2, the linear model is not appropriate. A first indication comes from the graph we made, which we repeat adding the regression line

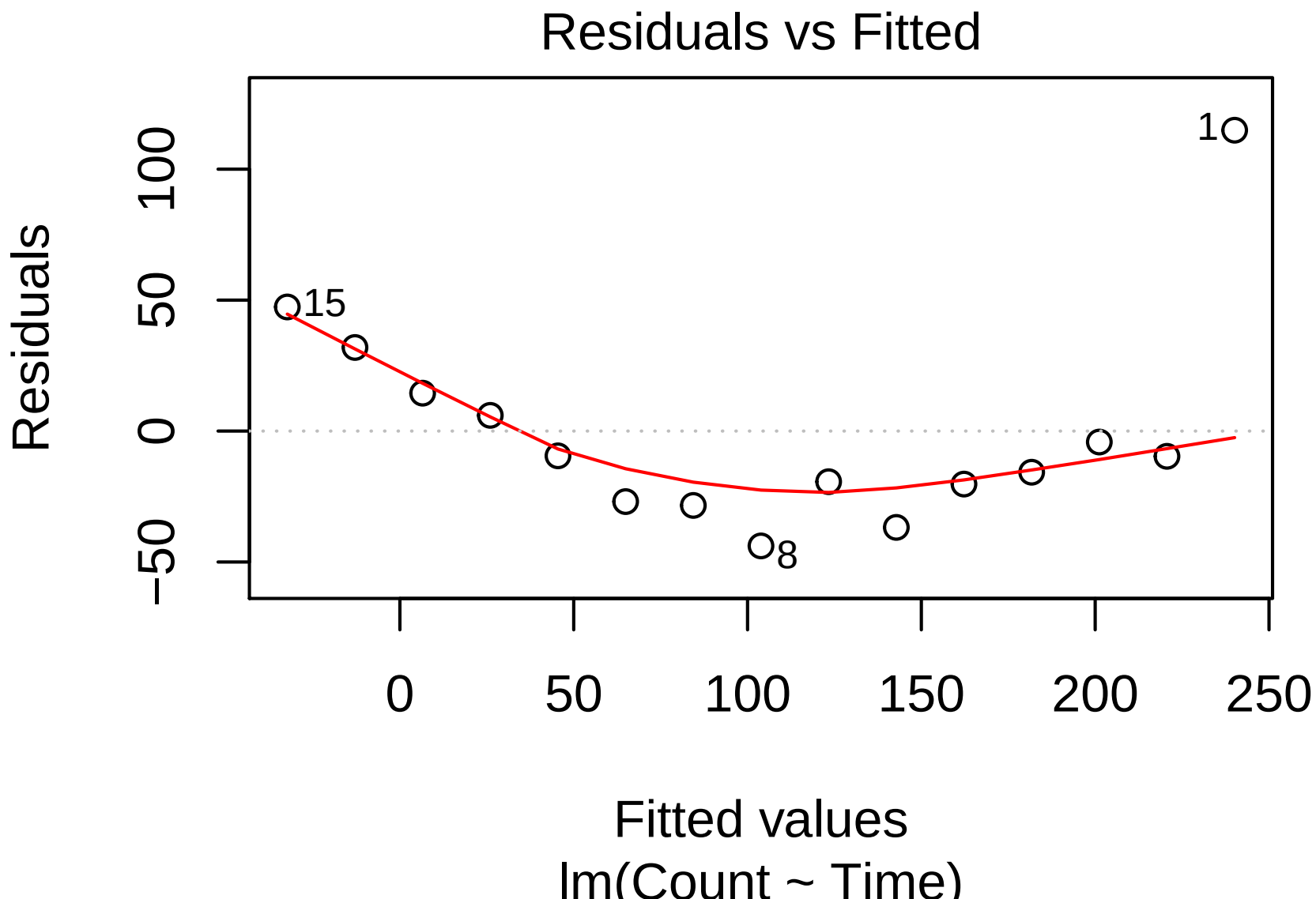
```
plot(Bacteria)
abline(fitbac)
```



Transformed Data

We find more evidence of this in the residuals graphs:

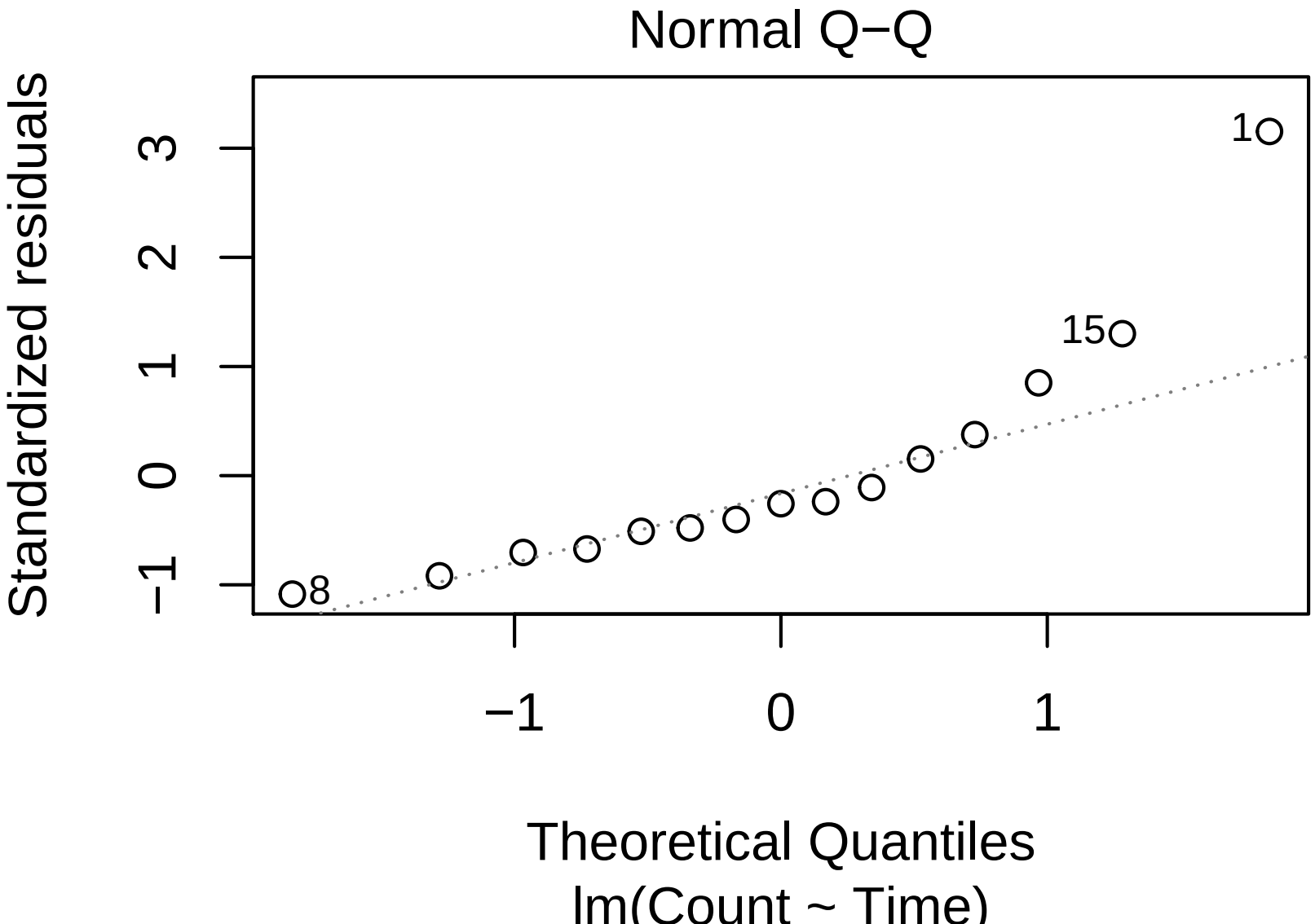
```
plot(fitbac, which=1)
```



Transformed Data

We find more evidence of this in the residuals graphs:

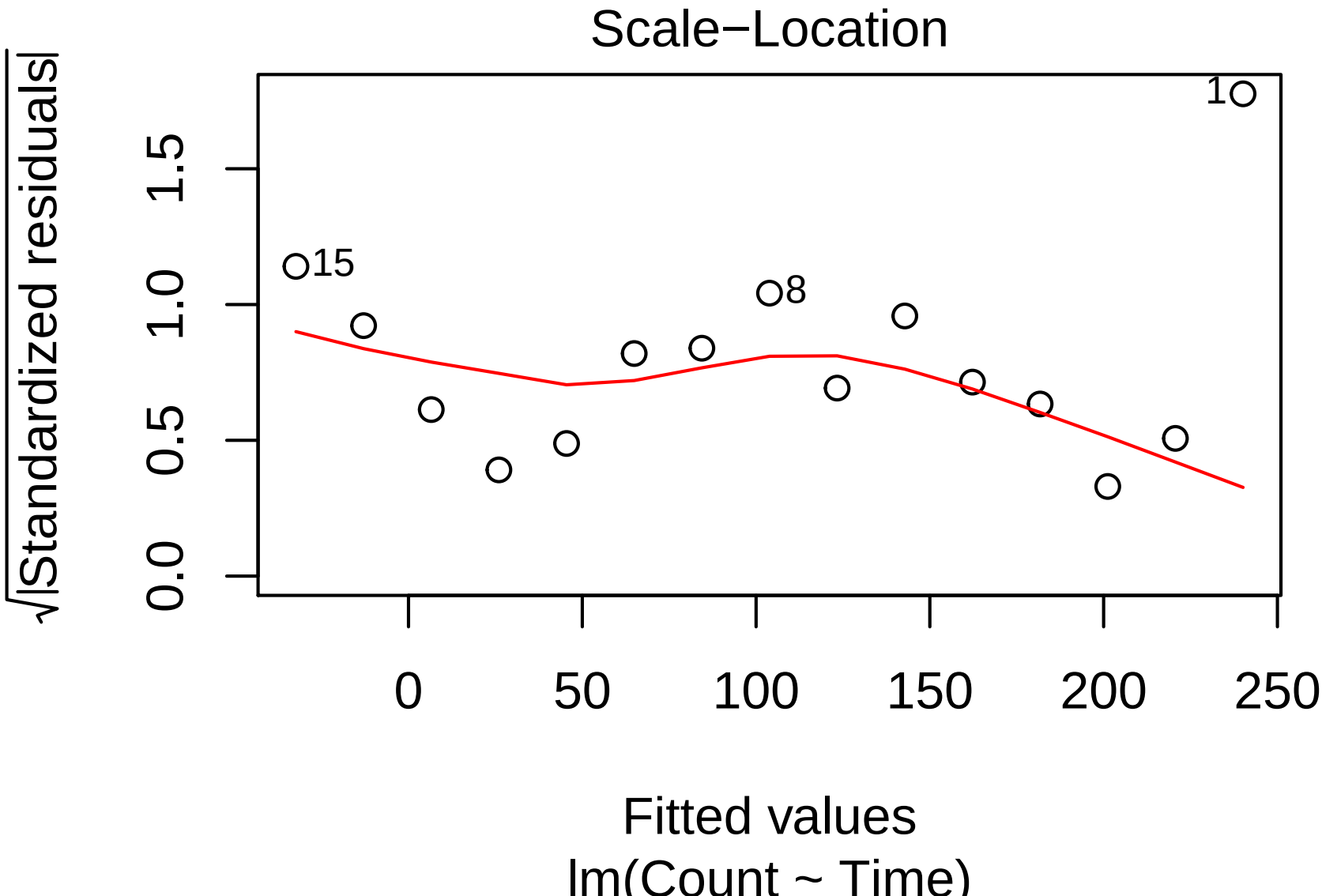
```
plot(fitbac, which=2)
```



Transformed Data

We find more evidence of this in the residuals graphs:

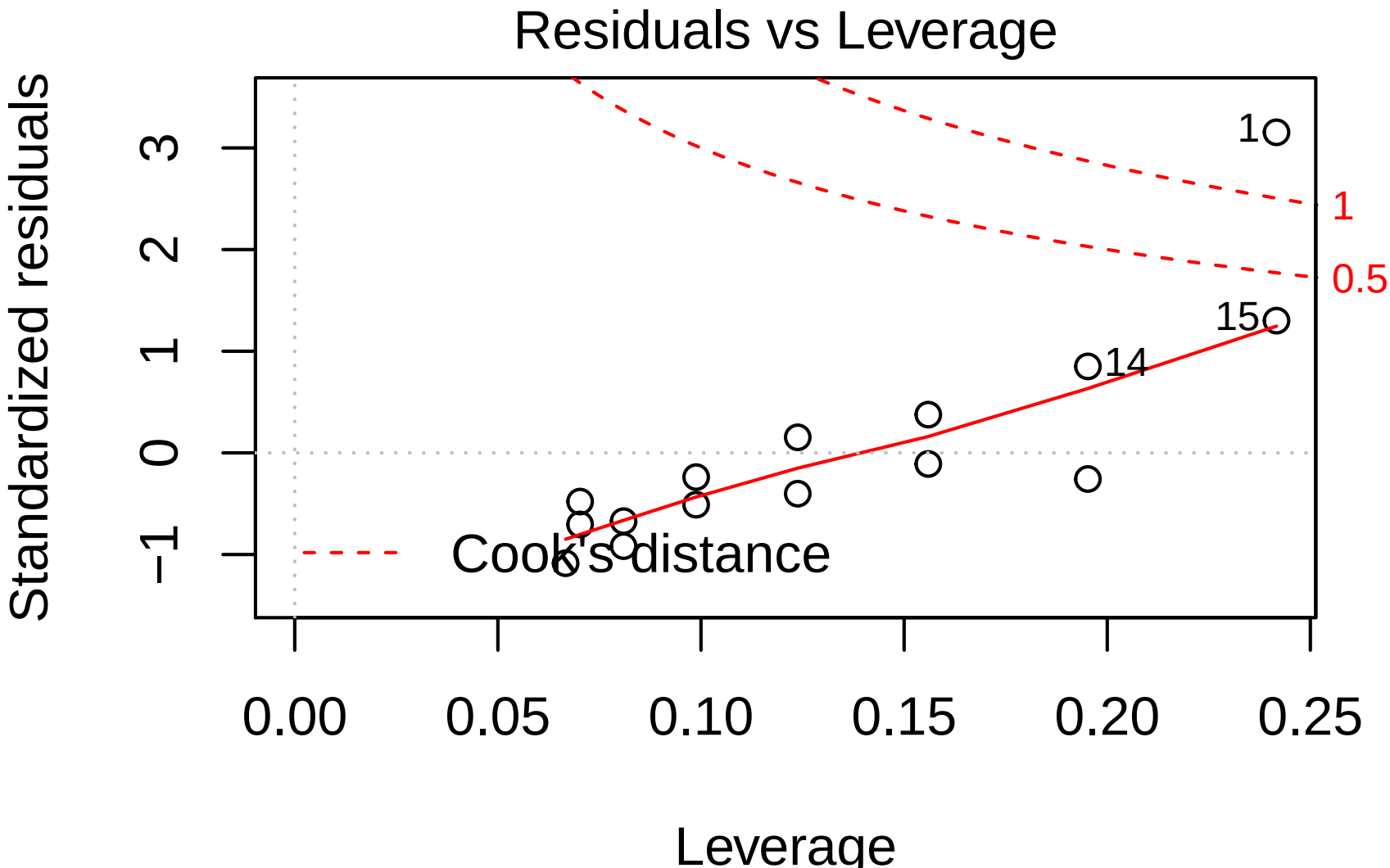
```
plot(fitbac, which=3)
```



Transformed Data

We find more evidence of this in the residuals graphs:

```
plot(fitbac, which=5)
```



Transformed Data

The first graph, residuals vs. fitted values, shows that this model does not explain all the relation existing between these two variables.

The quantile plot shows disagreement on the right tail of the distribution.

The third graph shows again that there is a structure in the residuals that has not been included in the model.

Finally, the last graph shows that there are some highly influential points in the regression.

Transformed Data

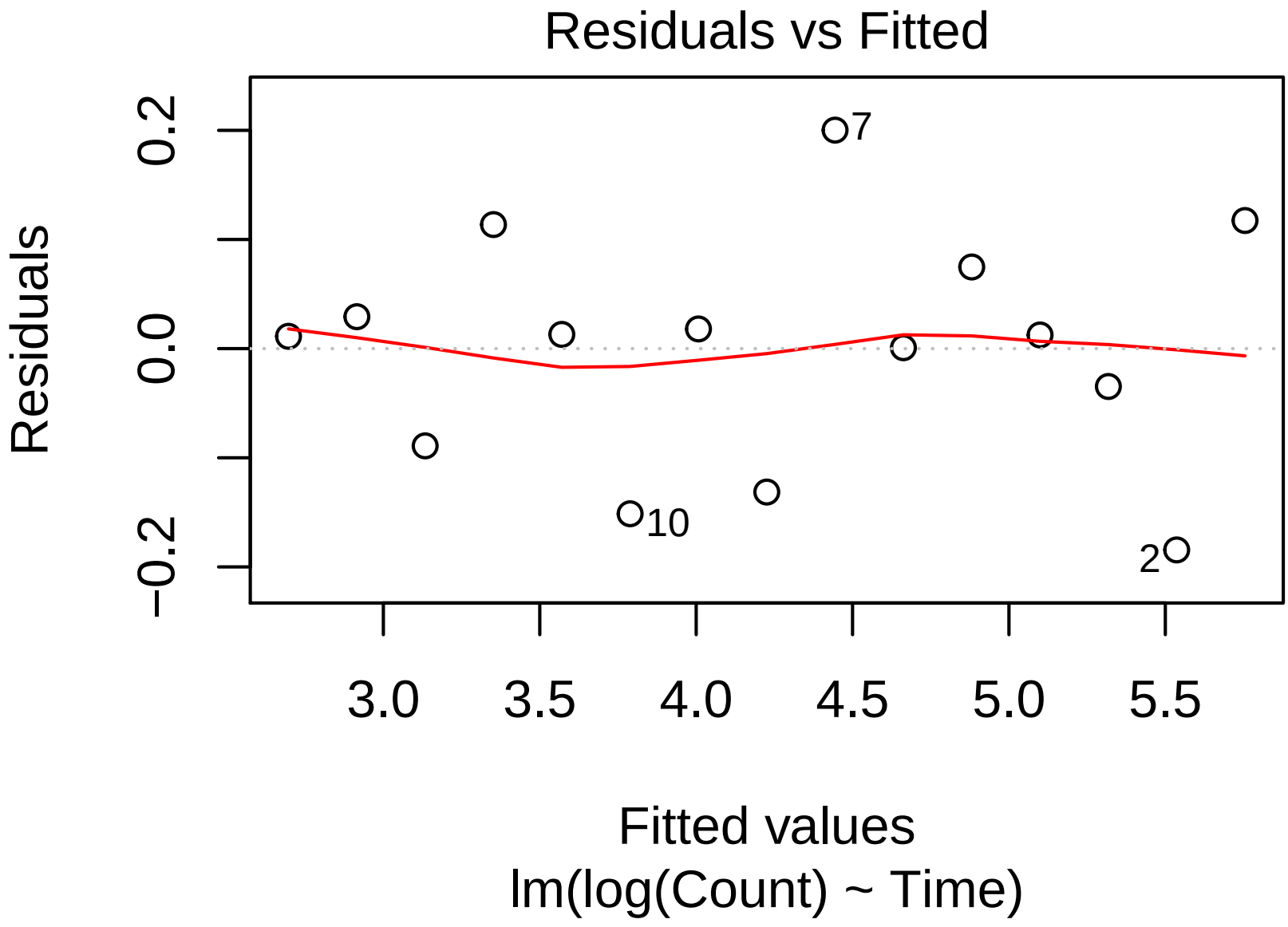
Let's look now at the model using a logarithmic transformation

```
fitlogbac <- lm(log(Count) ~ Time)
summary(fitlogbac)
```

```
##
## Call:
## lm(formula = log(Count) ~ Time)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18445 -0.06189  0.01253  0.05201  0.20021
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.973160   0.059778   99.92  < 2e-16 ***
## Time        -0.218425   0.006575  -33.22 5.86e-14 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.11 on 13 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.9875
## F-statistic:  1104 on 1 and 13 DF,  p-value: 5.86e-14
```

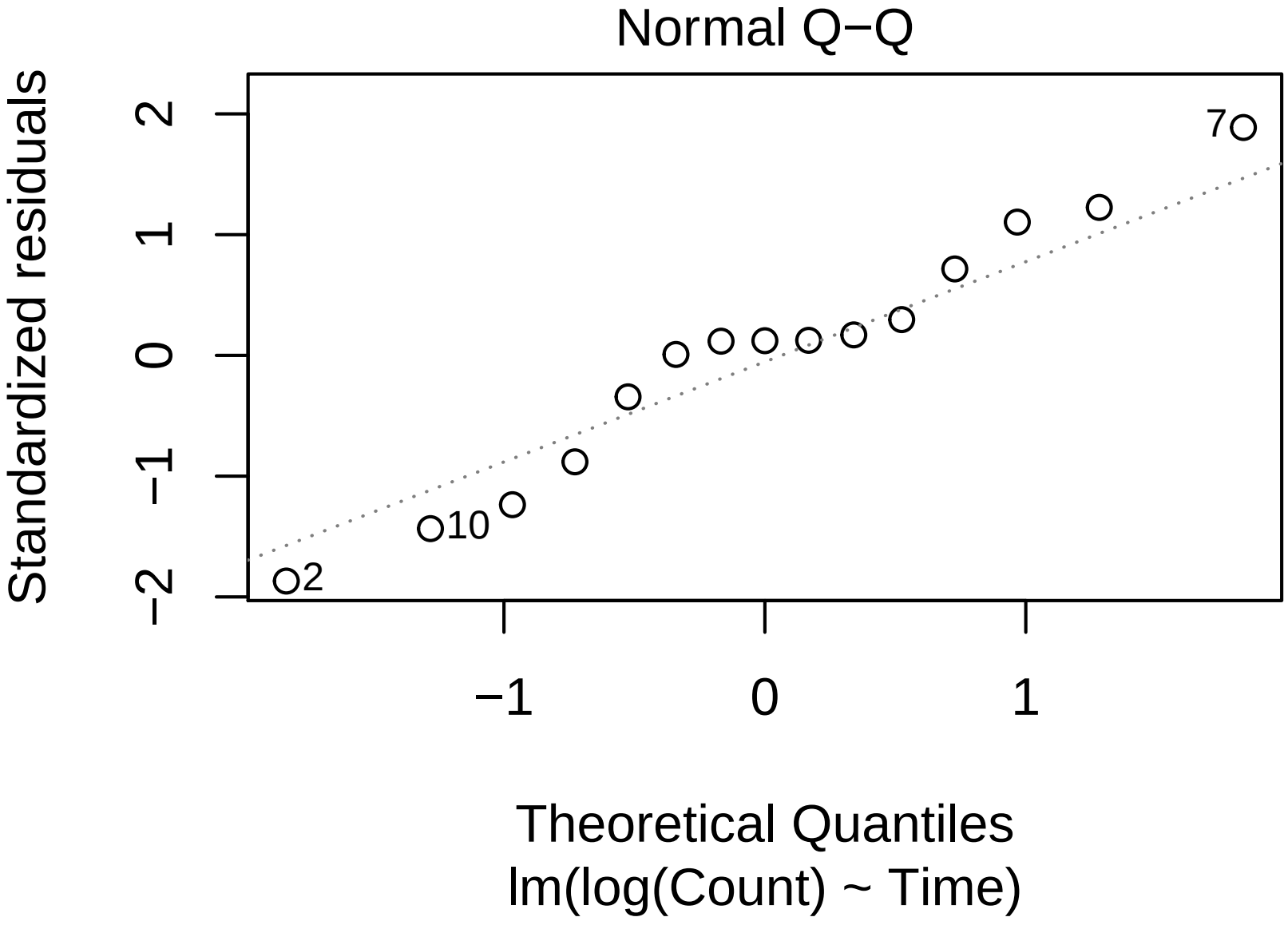
Transformed Data

```
plot(fitlogbac, which=1)
```



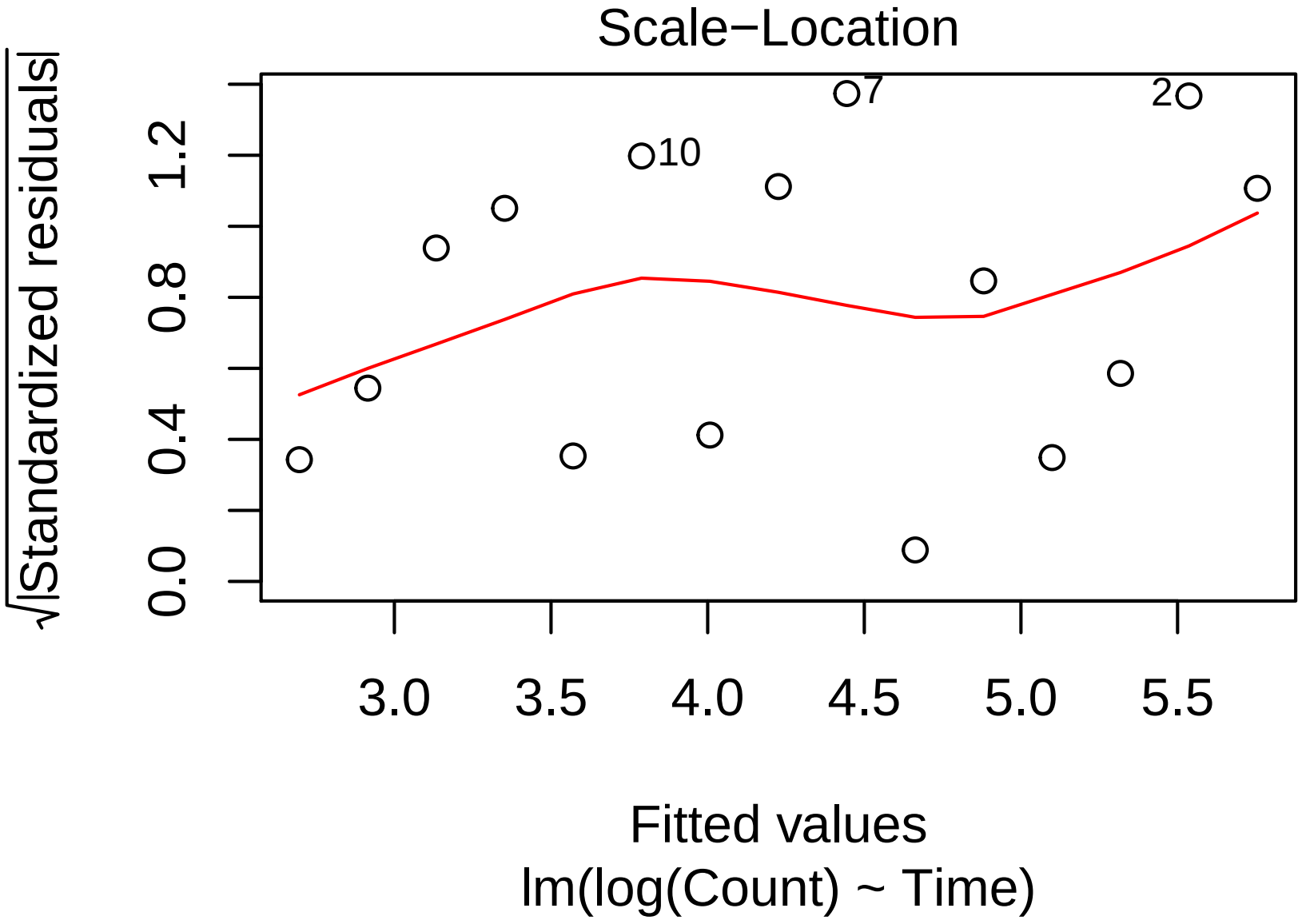
Transformed Data

```
plot(fitlogbac, which=2)
```



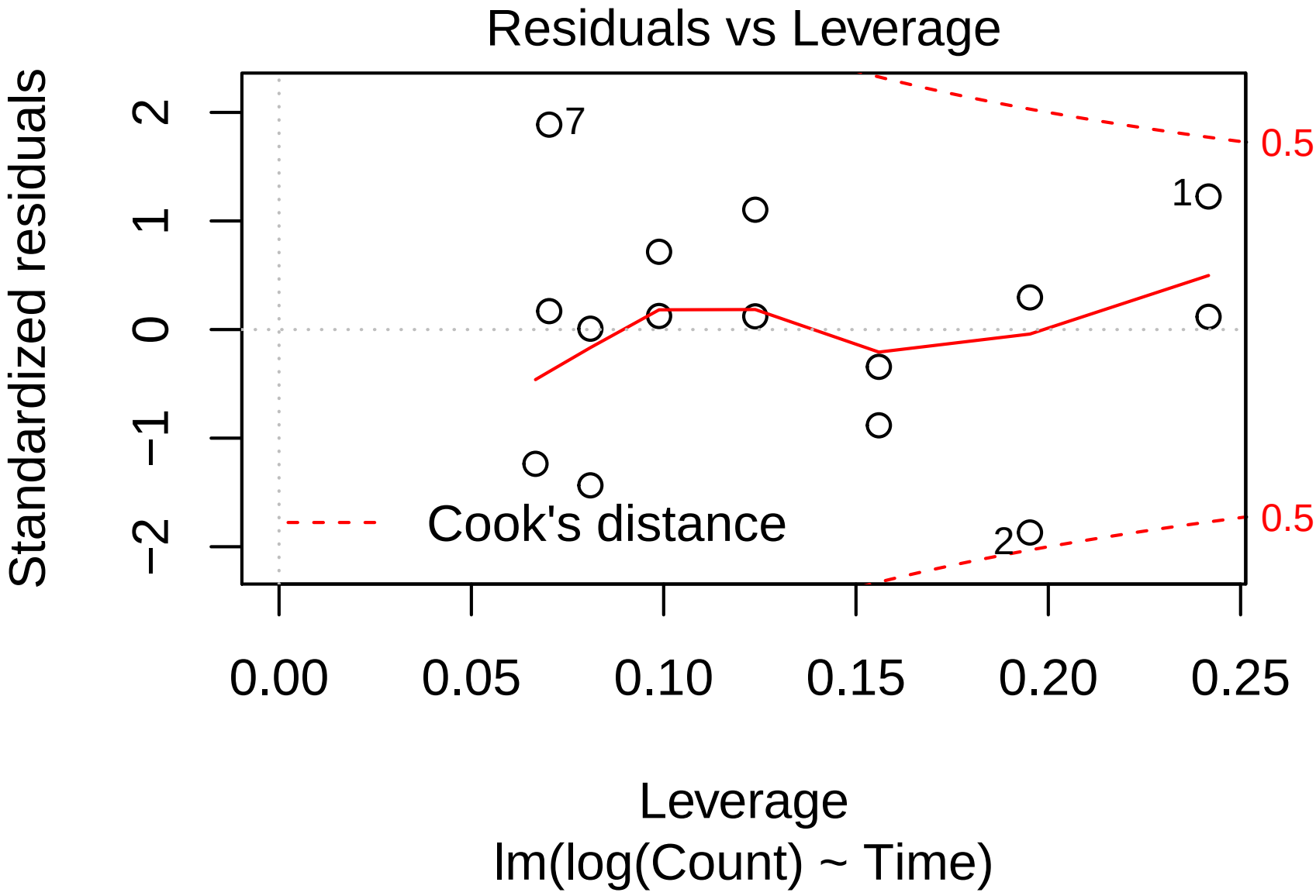
Transformed Data

```
plot(fitlogbac, which=3)
```



Transformed Data

```
plot(fitlogbac, which=5)
```



Transformed Data

We now see that the regression coefficients are significant, standard errors are reasonable, and the model explains about 98% of the variation in the data.

The residual graphs also show a considerably improved fit.

The residuals appear to be randomly distributed, the fit of the experimental data and model predictions is good, and the dispersion of the residuals has been considerably reduced.

The only graph that is not entirely satisfactory is the normal qq-plot, but this may be because we have little data.

Transformed Data

The linear model for $log(Count)$ is

$$\log(Count) = 5.97 - 0.219 \cdot t$$

where t is time, and the exponential model is

$$Count = e^{5.97 - 0.219 \cdot t} = 392.75 e^{-0.219 \cdot t}$$