Variance Stablization Transformation/ Regression

Yes

Stablizing the variance: for non-constant variance situation

R for the Case Example – Plutonium Measurement, p 141 of textbook#

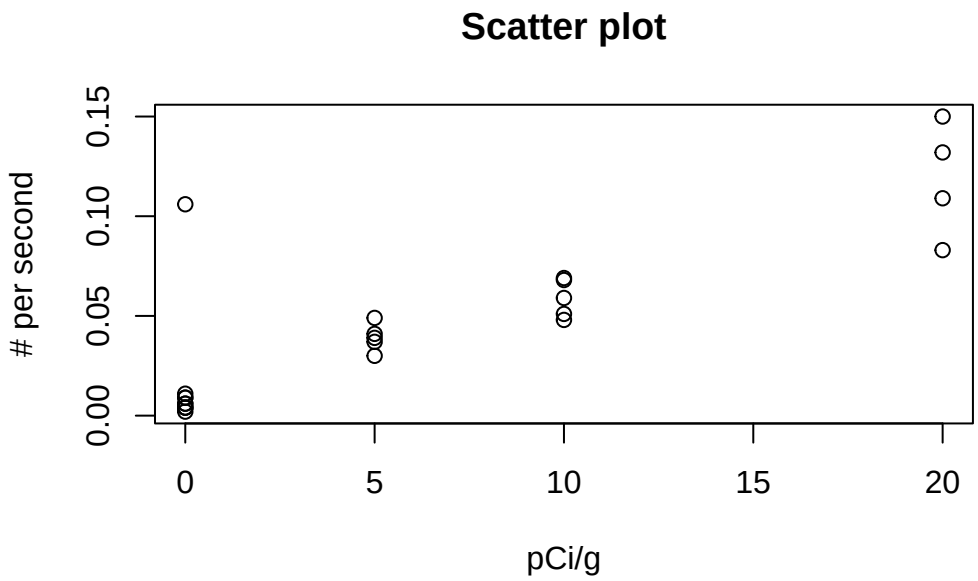
Background: Some environmental cleanup work requires that nuclear material, such as plutonium 238, be located and completely removed from a restoration site. When plutonium has become mixed with other materials in very small amounts, detecting its presence can be a difficult task. Even very small amounts can traced, however, because plutonium emits subatomic particles– alpha particles – that can be detected. Devices that are used to detect plutonium record the intensity of alpha particles strikes in counts per second (# per sec, let it be Y). The regression relationship between alpha counts per second (the response variable) and plutonium activity (the explanatory variable, let it be x) is then used to to estimate the activity of plutonium in the material under study. (This use of a regression relationship involves inverse prediction; see p.168 ~ 170 for details.)

Task: to estimate the regression relationship between alpha counts per second and plutonium activity.

Experiment: four standard rods conataining fixed, known level of plutonium activity:0.0, 5.0, 10.0 and 20.0 picocuries per gram (pCi/g) are exposed to the detection device from 4 to 10 times, and the rate of alpha strikes (counts per second) was recorded.

Data set : 24 data points

```
ch3ta10<-matrix(scan("CH03TA10.txt"),ncol=2, byrow=T) ;
# x <- ch3ta10[,2];
# y <- ch3ta10[,1];
dum <- data.frame(x=ch3ta10[,2], y=ch3ta10[,1])
attach(dum)
# Scatter plot to see if a linear relation suitable....
# windows()
plot(x,y,xlab="pCi/g", ylab="# per second", main="Scatter plot");
```




```
# Notice that, as expected, the strike rate tends to increase with the plutonium
# activity level.....  except one strange data point... (which one?)
# Note, also, a nonzero strike rate for the standard rod with zero plutonium level
# (possible reason?)

# dum <- data.frame(x=ch3ta10[,2], y=ch3ta10[,1])
# attach(dum)
# rm(x,y)
```

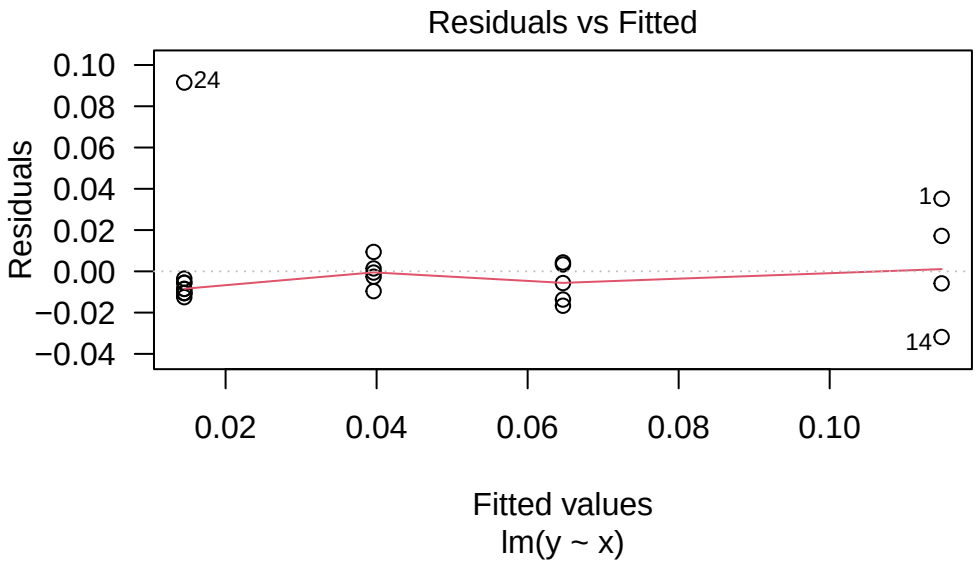
```r
# simple regression: y reg. on x...
fm <- lm(y~x, dum)
summary(fm)
```

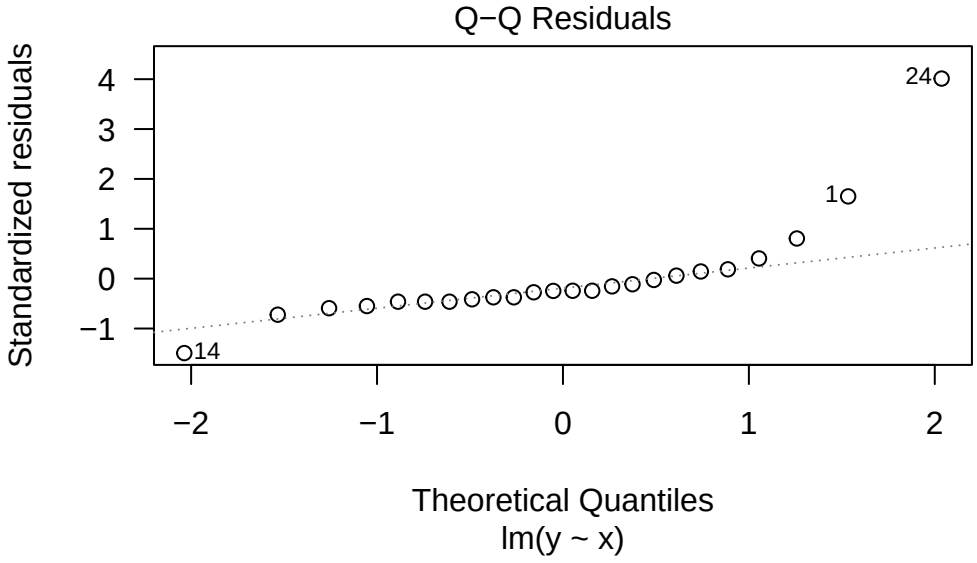
```
Call:
lm(formula = y ~ x, data = dum)

Residuals:
      Min        1Q    Median        3Q       Max
-0.031826 -0.010529 -0.005603  0.001878  0.091471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0145294  0.0065264   2.226   0.0366 *
x           0.0050148  0.0006778   7.398 2.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02371 on 22 degrees of freedom
Multiple R-squared:  0.7133,    Adjusted R-squared:  0.7003
F-statistic: 54.74 on 1 and 22 DF,  p-value: 2.107e-07
```

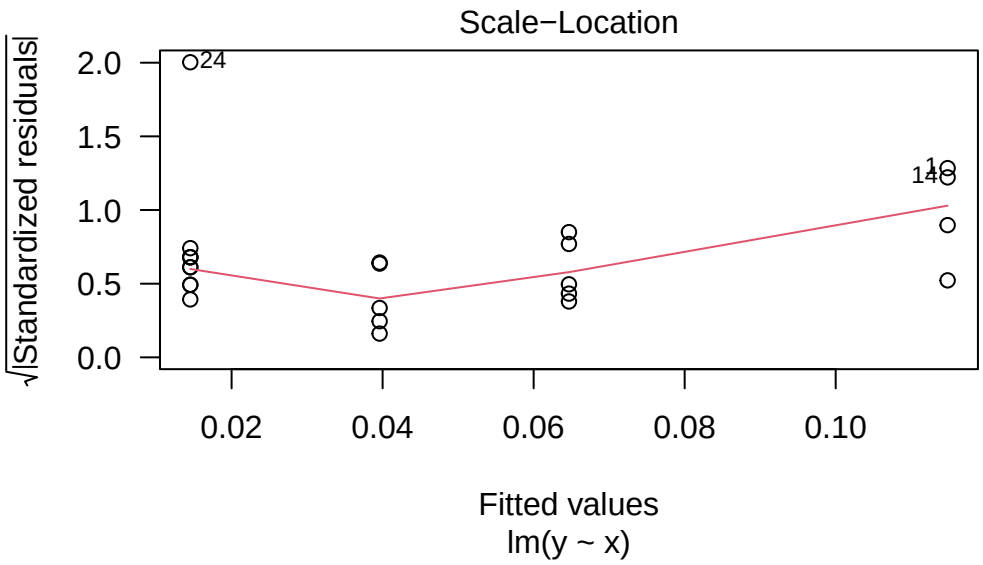
```r
anova(fm)
```

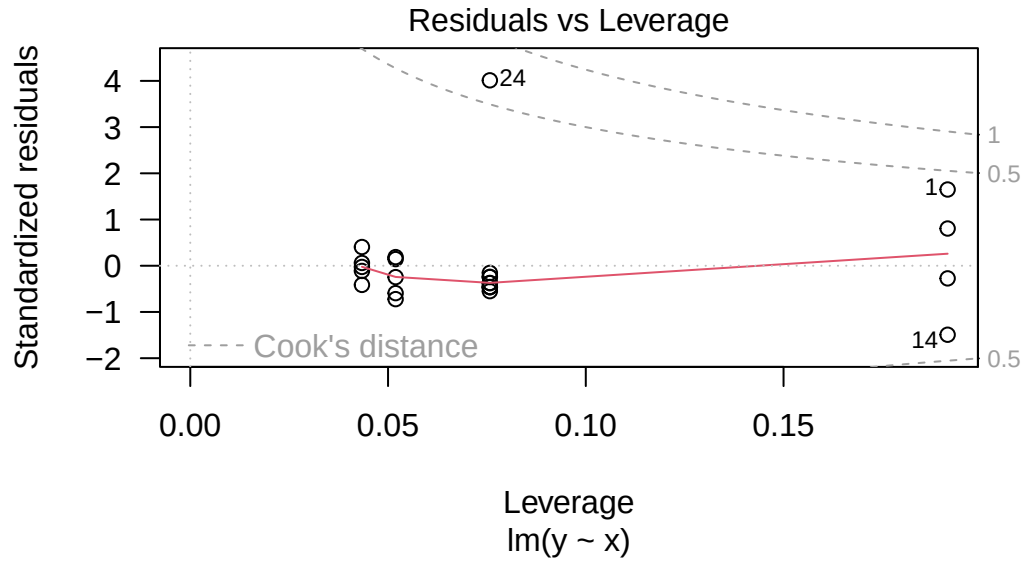
```
Analysis of Variance Table

Response: y
          Df   Sum Sq   Mean Sq F value    Pr(>F)
x          1 0.030780 0.0307805  54.737 2.107e-07 ***
Residuals 22 0.012371 0.0005623
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(fm, las=1);       # Note: R gives four basic plots for diagnostics.... #
```


Residuals vs Fitted
Residuals
0.10
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
24
1
14
0.02
0.04
0.06
0.08
0.10
Fitted values
lm(y ~ x)


Q−Q Residuals
Standardized residuals
4
3
2
1
0
−1
24
1
14
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)






```
# case 24 is an outlier.  An examination of laboratory record revealed that
# the experimental conditions were not properly maintained for that case.
# It is decided, then, to discard case 24 in the analysis....
```

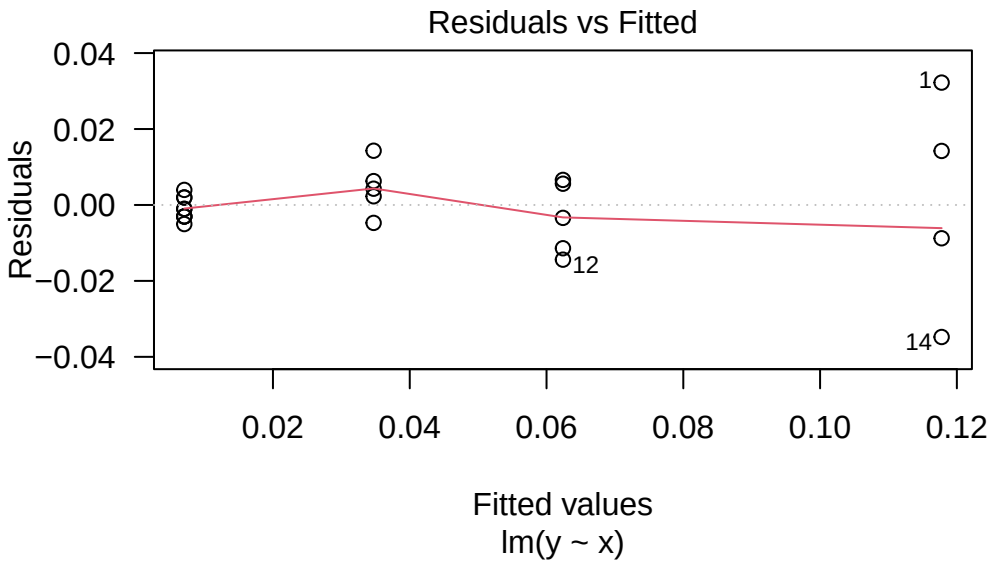
```r
fm0 <- lm(y~x, dum[-24, ])       # remove case 24 from dum
summary(fm0)
```

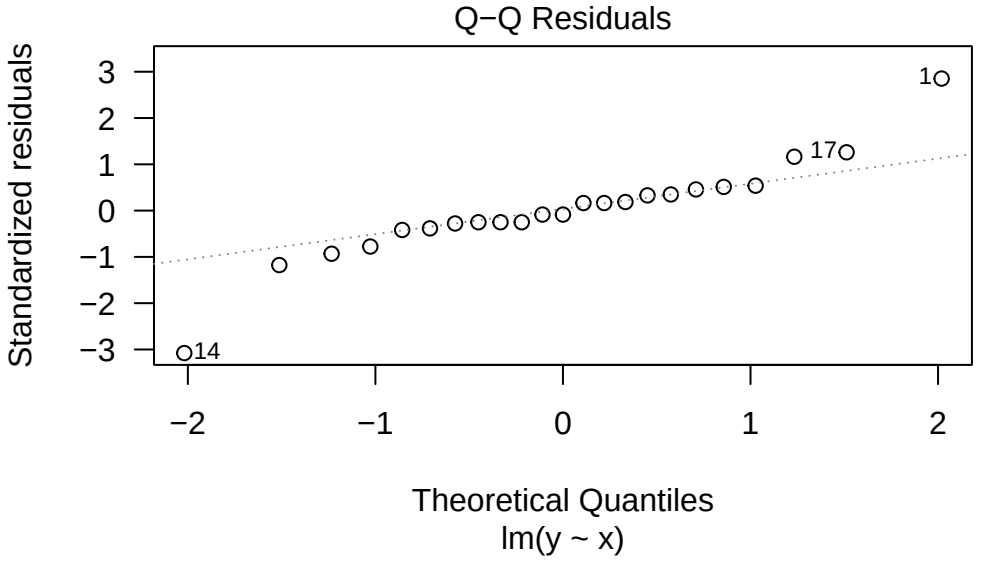
```
Call:
lm(formula = y ~ x, data = dum[-24, ])

Residuals:
      Min        1Q    Median        3Q       Max
-0.034773 -0.004061 -0.001033  0.004939  0.032227

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0070331  0.0035988   1.954   0.0641 .
x           0.0055370  0.0003659  15.133 9.08e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01257 on 21 degrees of freedom
Multiple R-squared:  0.916, Adjusted R-squared:  0.912
F-statistic:   229 on 1 and 21 DF,  p-value: 9.077e-13
```

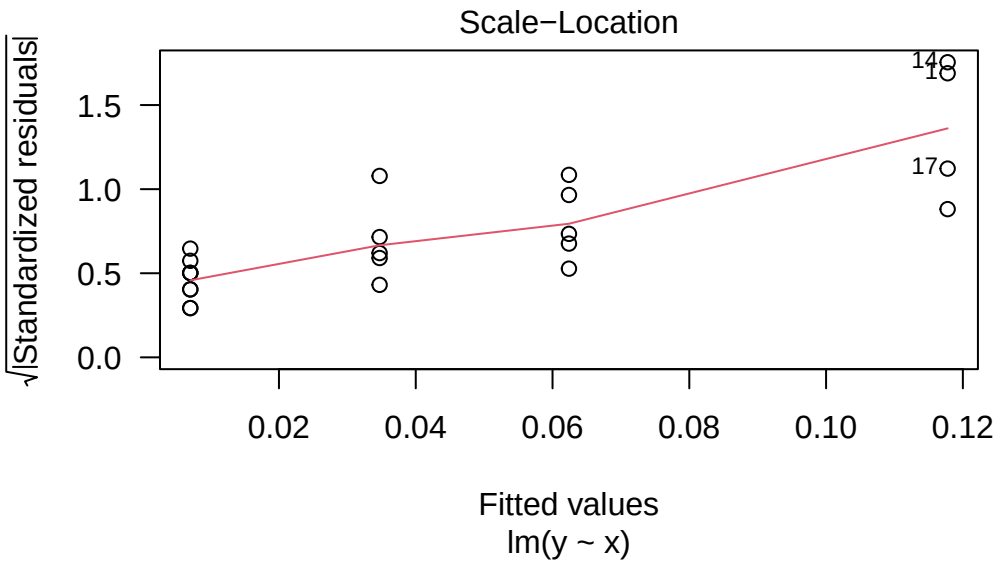
```r
anova(fm0)
```

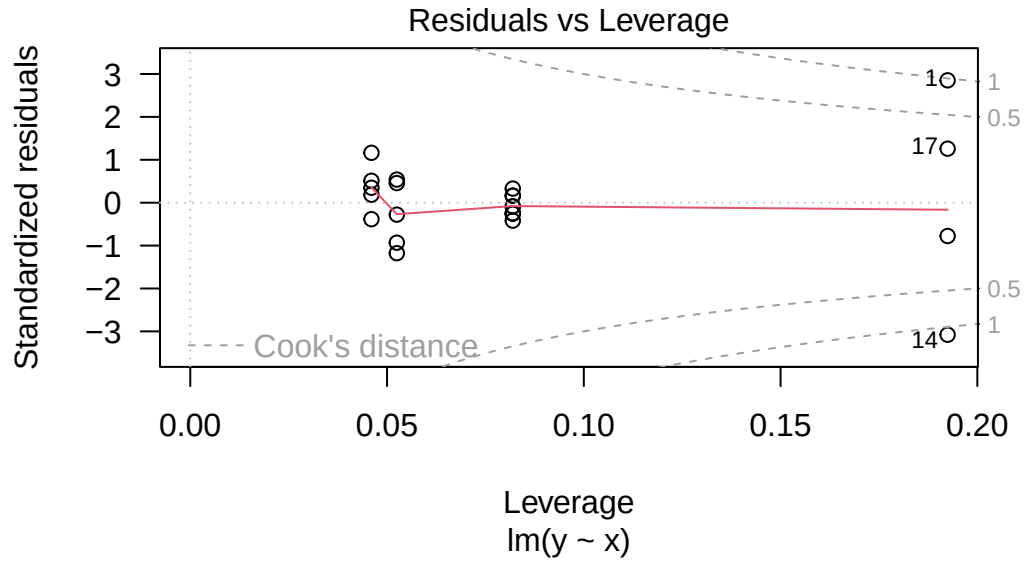
```
Analysis of Variance Table

Response: y
          Df   Sum Sq  Mean Sq F value    Pr(>F)
x          1 0.036190 0.036190     229 9.077e-13 ***
Residuals 21 0.003319 0.000158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(fm0, las=1);     # diagnostics
```


Residuals vs Fitted
Residuals
0.04
0.02
0.00
−0.02
−0.04
1
12
14
0.02
0.04
0.06
0.08
0.10
0.12
Fitted values
lm(y ~ x)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
−3
1
17
14
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)






```
# F-value=229 on 1, 21 df's, p-value =0, so there is a significant regression
# relationship.
# However, error variance appears to increase with the plutonium activity level.
# Normal Q-Q plot suggests nonnormality (heavy tail)... but probably due to
```

```
# the unequal error variance....
# To get X^2_BP
SQe<-resid(fm0)^2          # e^2
dum1<-data.frame(x[-24],SQe)
attach(dum1)
```

```
The following object is masked _by_ .GlobalEnv:

    SQe
```

```
X2BP<-(sum((fitted(lm(SQe~x[-24],dum1))-mean(SQe)  )^2)/2)/(sum(resid(fm0)^2)
/(length(x[-24])))^2
X2BP
```

```
[1] 23.32601
```

```
X2BP>qchisq(.95,1)       #?, if yes, reject the constant variance null hypothesis
```

```
[1] TRUE
```

```
#or calculate the p-value of the test and reject if p-value < alpha
1-pchisq(X2BP,1)
```

```
[1] 1.36738e-06
```

```
#Test decision:
#Note: X^2_BP= 23.29 > 3.84   (or p-value < 0.05), that is the
# Breusch-Pagan test rejects at level 0.05 the equal variance null hypothesis.

# Try to transform y, then....   take sqrt(y) be our response variable, say...
sfm0 <- update(fm0,sqrt(.)~.)   # a very useful R command.....
summary(sfm0)
```

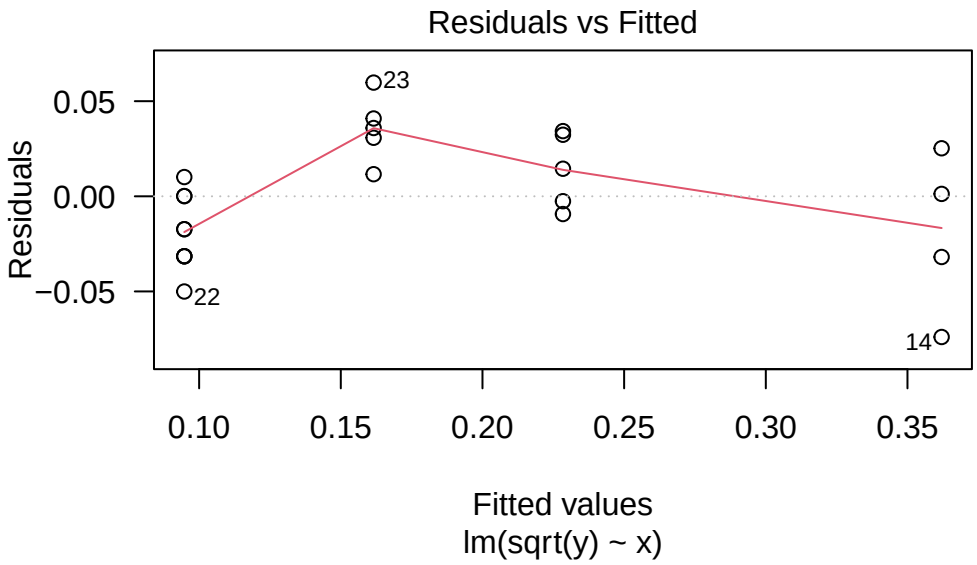
```
Call:
lm(formula = sqrt(y) ~ x, data = dum[-24, ])

Residuals:
      Min        1Q    Median        3Q       Max
```

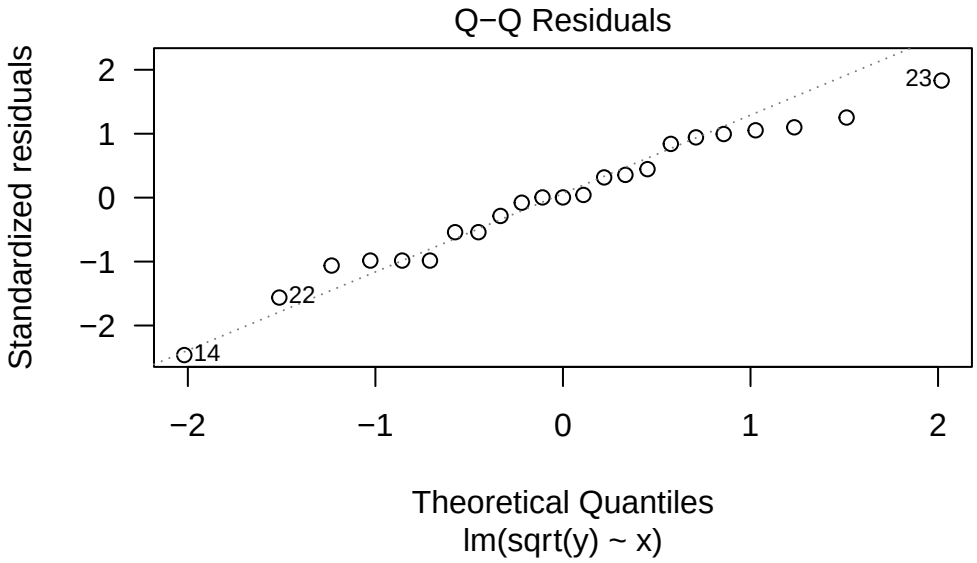
```
-0.073958 -0.024407  0.000109  0.028007  0.059776

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0947596  0.0095668   9.905 2.29e-09 ***
x           0.0133648  0.0009727  13.740 5.77e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03342 on 21 degrees of freedom
Multiple R-squared:  0.8999,    Adjusted R-squared:  0.8951
F-statistic: 188.8 on 1 and 21 DF,  p-value: 5.767e-12
```

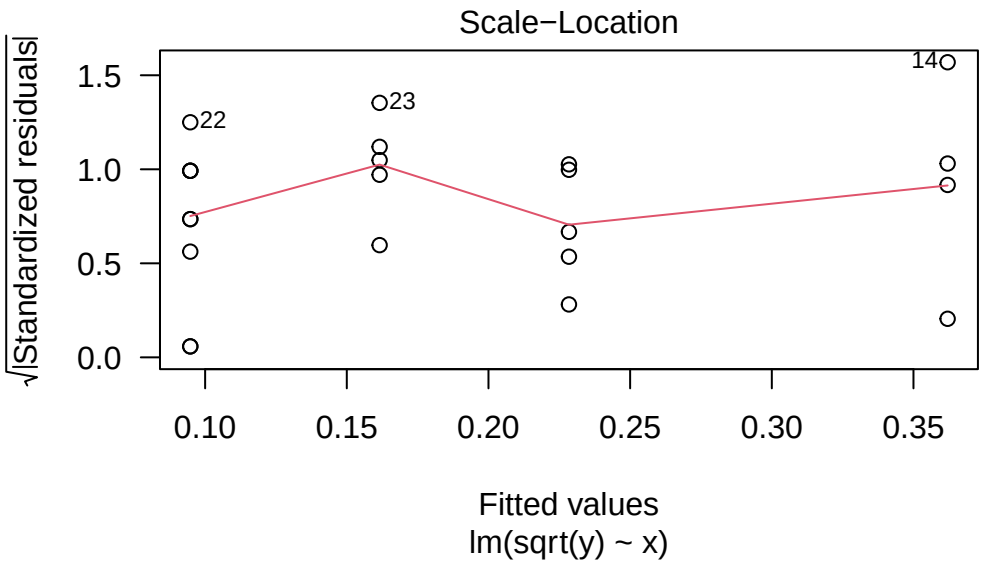
```r
anova(sfm0)
```

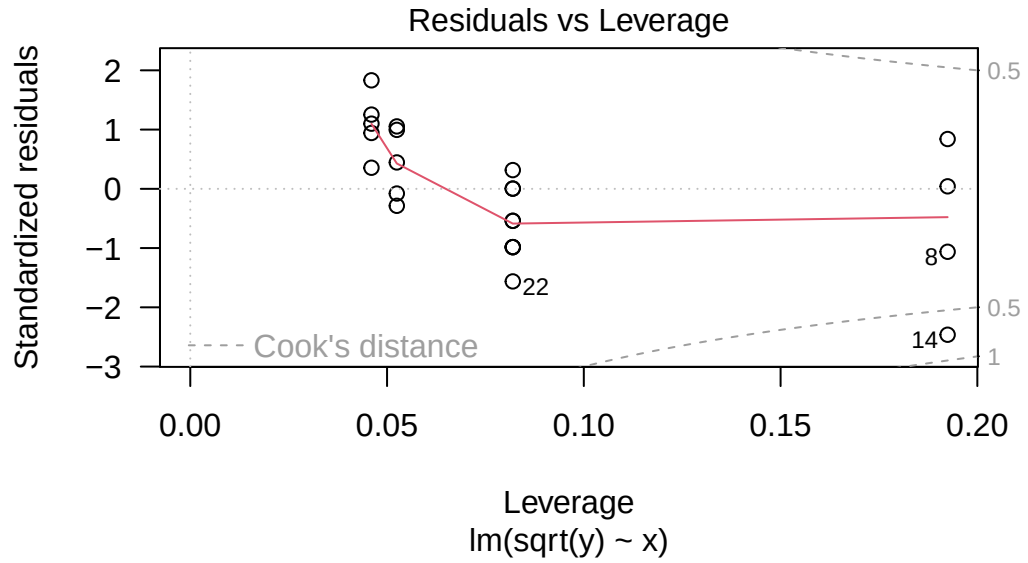
```
Analysis of Variance Table

Response: sqrt(y)
          Df   Sum Sq  Mean Sq F value    Pr(>F)
x          1 0.210847 0.210847   188.8 5.767e-12 ***
Residuals 21 0.023453 0.001117
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(sfm0, las=1)      # diagnostics
```


Residuals vs Fitted
Residuals
0.05
0.00
−0.05
23
22
14
0.10
0.15
0.20
0.25
0.30
0.35
Fitted values
lm(sqrt(y) ~ x)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
23
22
14
−2
−1
0
1
2
Theoretical Quantiles
lm(sqrt(y) ~ x)






```
# Normal Q-Q plot is now roughly a straight line
# but residuals v.s. y_hat: not a horizontal band... : suggests a nonlinear relation
# of course....  since y and x are related linearly... sqrt(y) and x, then,....
```

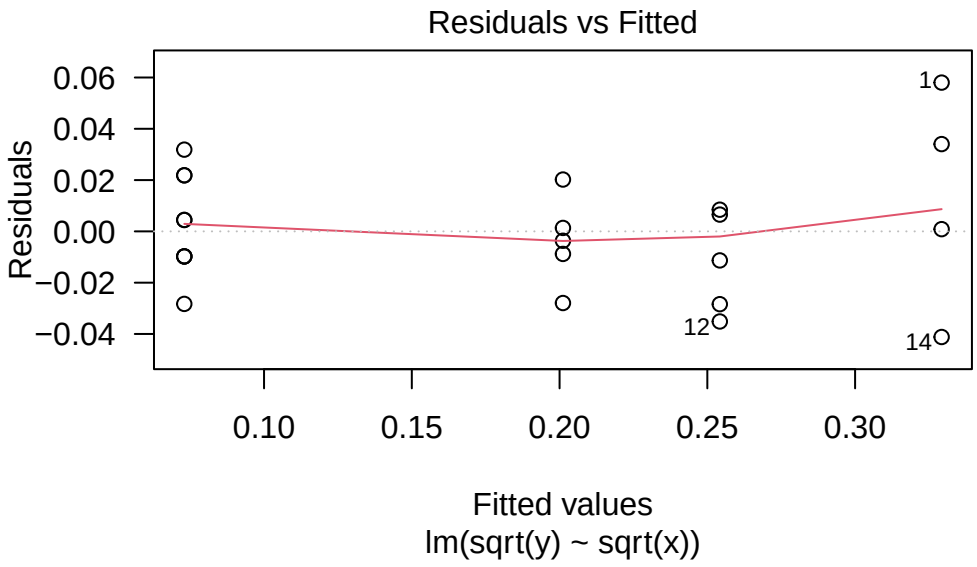
```r
# take sqrt(x) be the predictor for sqrt(y), then...
sfm0s <- update(sfm0, .~sqrt(.))      # really take a note on this command.... #
summary(sfm0s)
```

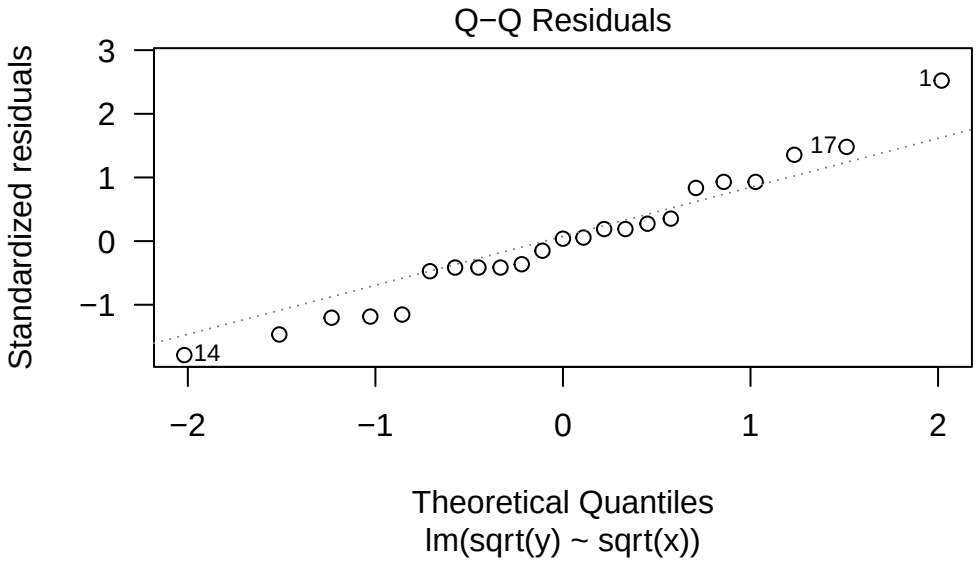
```
Call:
lm(formula = sqrt(y) ~ sqrt(x), data = dum[-24, ])

Residuals:
      Min        1Q    Median        3Q       Max
-0.041186 -0.010541  0.000868  0.014336  0.058015

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.073006   0.007831   9.323 6.51e-09 ***
sqrt(x)     0.057305   0.003016  18.998 1.05e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02477 on 21 degrees of freedom
Multiple R-squared:  0.945, Adjusted R-squared:  0.9424
F-statistic: 360.9 on 1 and 21 DF,  p-value: 1.046e-14
```

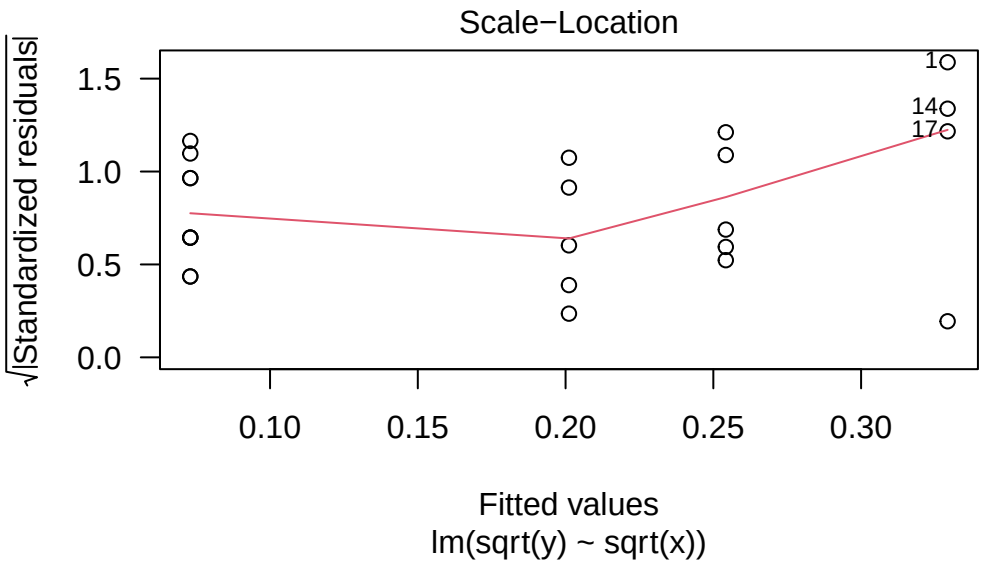
```r
anova(sfm0s)
```

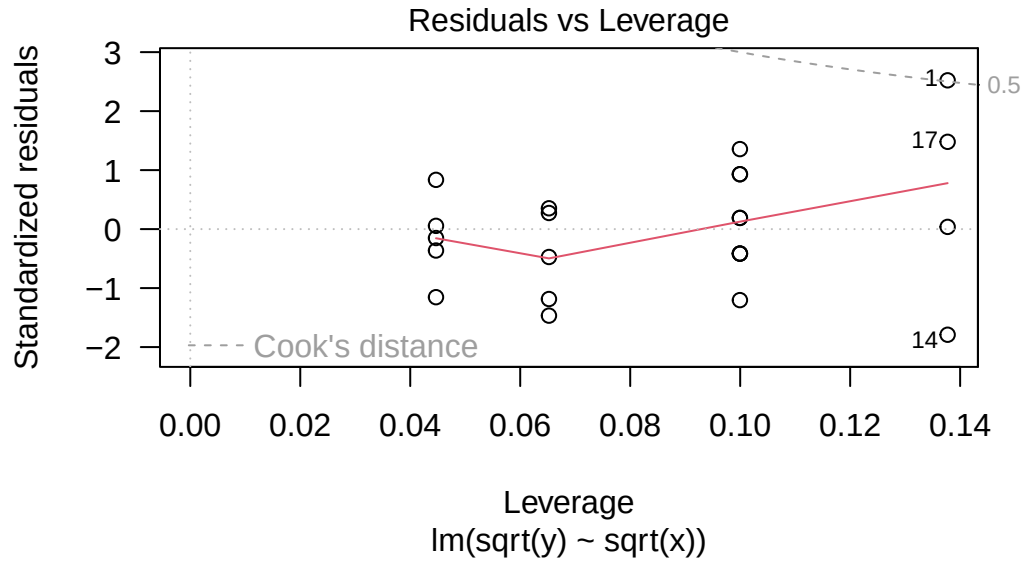
```
Analysis of Variance Table

Response: sqrt(y)
          Df   Sum Sq  Mean Sq F value    Pr(>F)
sqrt(x)    1 0.221416 0.221416  360.92 1.046e-14 ***
Residuals 21 0.012883 0.000613
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(sfm0s, las=1);   # diagnostics
```


Residuals vs Fitted
Residuals
0.06
0.04
0.02
0.00
−0.02
−0.04
1
12
14
0.10
0.15
0.20
0.25
0.30
Fitted values
lm(sqrt(y) ~ sqrt(x))


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
1
17
14
−2
−1
0
1
2
Theoretical Quantiles
lm(sqrt(y) ~ sqrt(x))






```
# A satisfactory linear fit results..... ;>
# Residual plots does suggest, though, that some nonconstancy of the error
# variance may still remain; but if so, it does not appear to be substantial.
# since X^2_BP=3.85 w/ p-value=0.05,  supporting the conclusion from the residual plot
```

```
SQe<-resid(sfm0s)^2          # e^2
dum2<-data.frame(sqrt(x)[-24],SQe)
attach(dum2)
```

```
The following object is masked _by_ .GlobalEnv:

    SQe

The following object is masked from dum1:

    SQe
```

```
X2BP<-(sum((fitted(lm(SQe~sqrt(x)[-24],dum2))-mean(SQe)  )^2)/2)/(sum(resid(sfm0s)^2)
/(length(x[-24])))^2
X2BP
```

```
[1] 3.852911
```

```
X2BP>qchisq(.95,1)       #?, if yes, reject the constant variance null hypothesis
```

```
[1] TRUE
```

```
#or calculate the p-value of the test and reject if p-value < alpha
1-pchisq(X2BP,1)
```

```
[1] 0.04965974
```

```
# So the nonconstancy of the error variance is not substantial.... say, "satisfactory" then

# The final fitted line: sqrt(y) = 0.0730 + 0.0573 sqrt(x) with R^2=0.945
# F-value = 360.9 with p-value=0: fitted line is statistically significant.

# With this data set: the estimated regression relationship between
# alpha counts per second and plutonium activity is
# sqrt(alpha counts per second) = 0.073 + 0.0573 * sqrt(plutonium activity level)
# The estimated reg. relationship is statistically significant.
```