Regression

Anatoly Dryga

2/17/2016

lm() function For Regression

Explanatory variable(s) is continuous.

Assumptions:

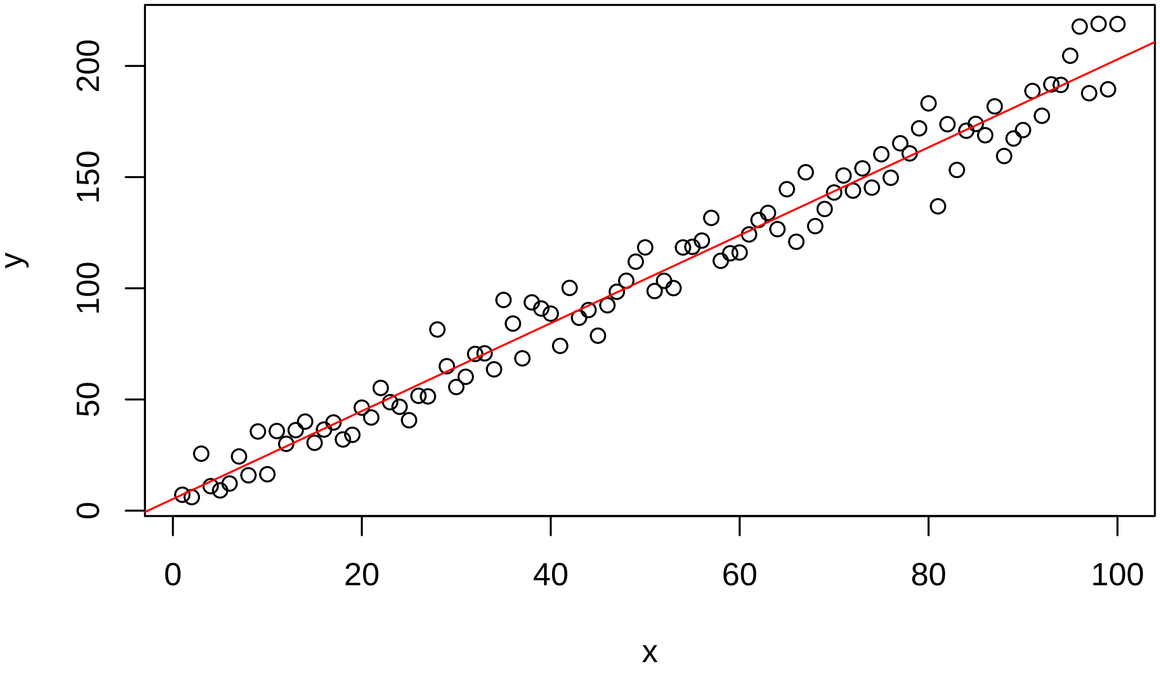
- errors are normally distributed
- variances are constant
- the explanatory variable is measured without error

Model

$$y = a + b \cdot x$$

The simplest case

```
x <- seq(1:100)
y <- 4 + 2*x + 10*rnorm(length(x))
plot(x, y)
df <- as.data.frame(list(y=y, x=x))
lin <- lm(y ~ x, df)
abline(lin, col="red")
```



```r
summary(lin)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.5084  -6.3530  -0.2458   6.3823  22.6758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.24187    1.96204   2.672  0.00884 **
## x            1.97699    0.03373  58.611  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.737 on 98 degrees of freedom
## Multiple R-squared:  0.9723, Adjusted R-squared:  0.972
## F-statistic:  3435 on 1 and 98 DF,  p-value: < 2.2e-16
```

Variance

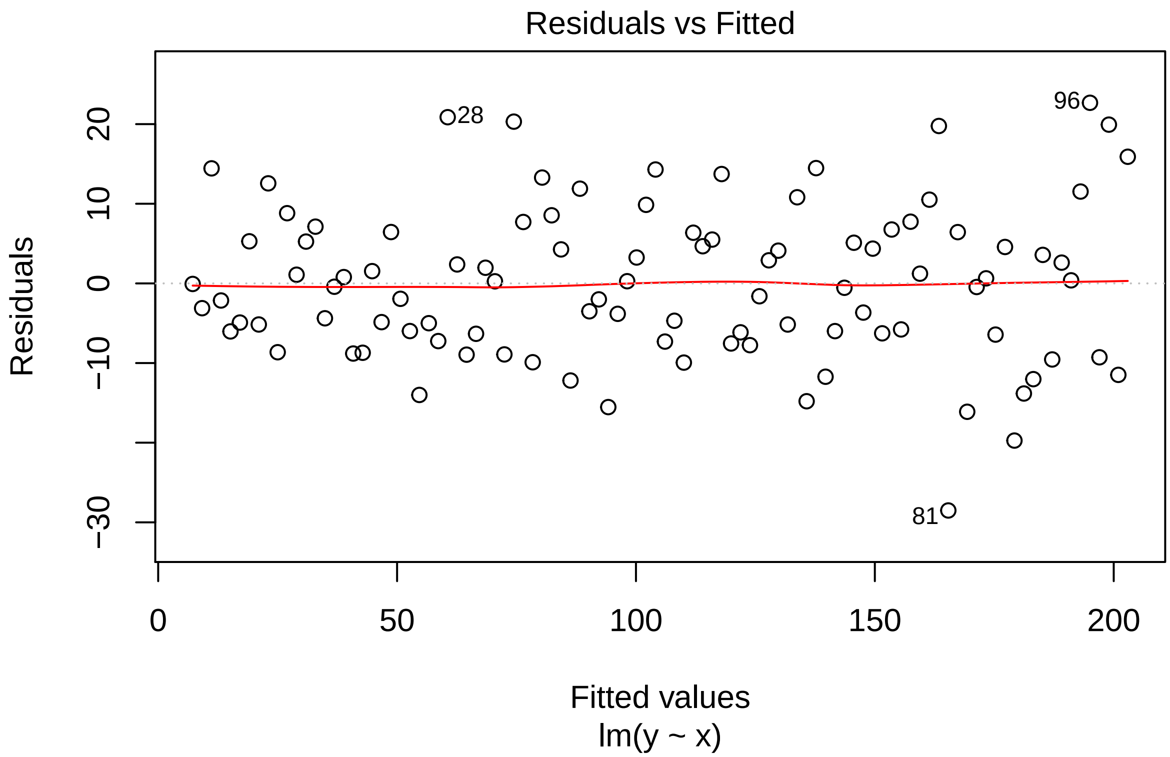
Sum of squares can be partitioned:

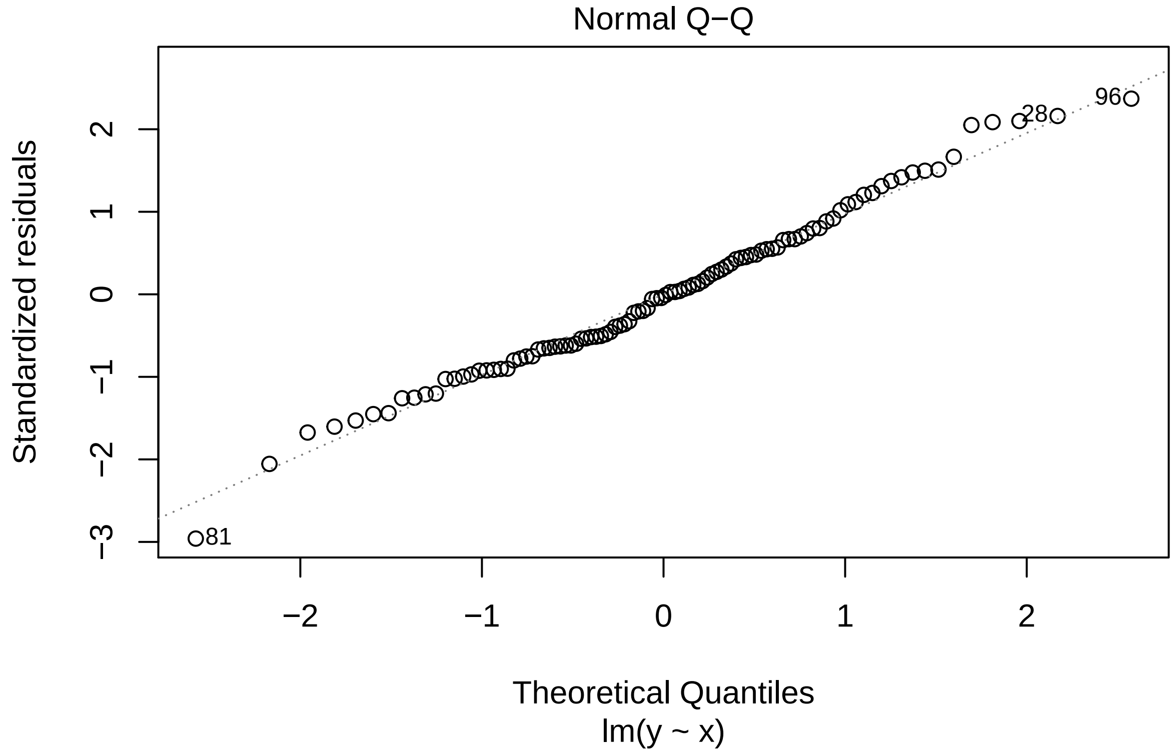
$$SST = SSR(regression) + SSE(residual)$$

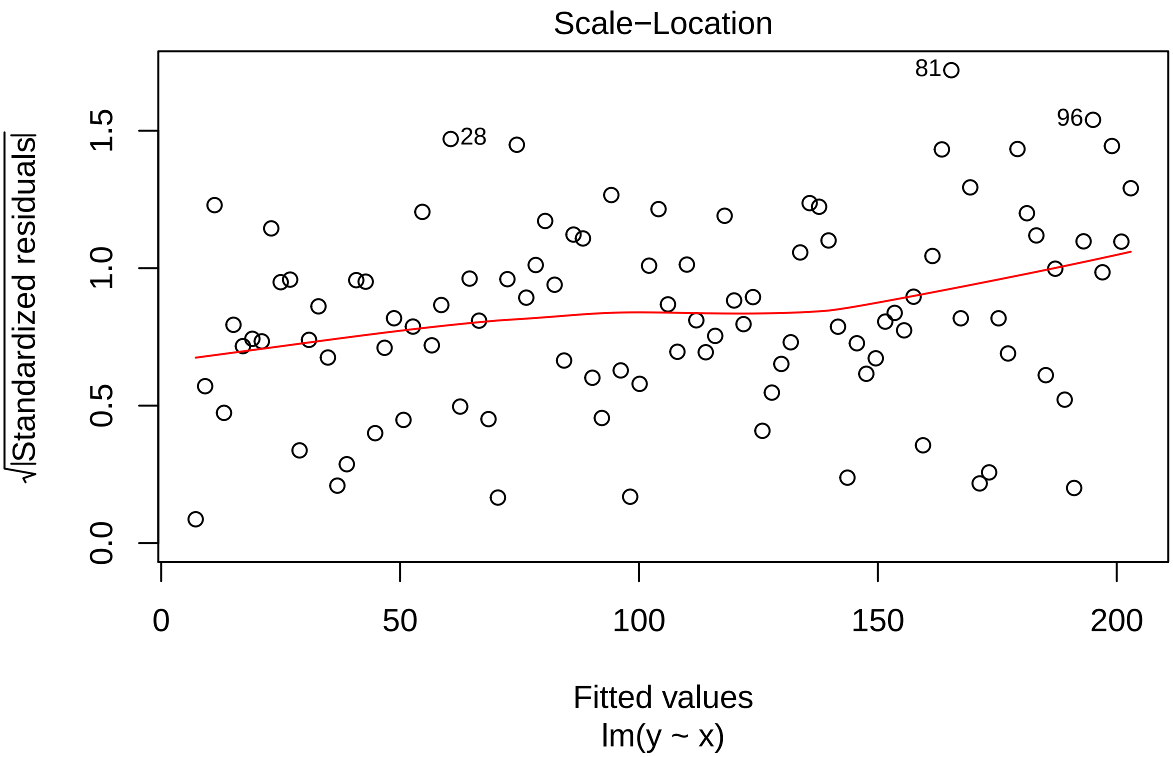
Total variance is explained by regression line and also has unexplained component.

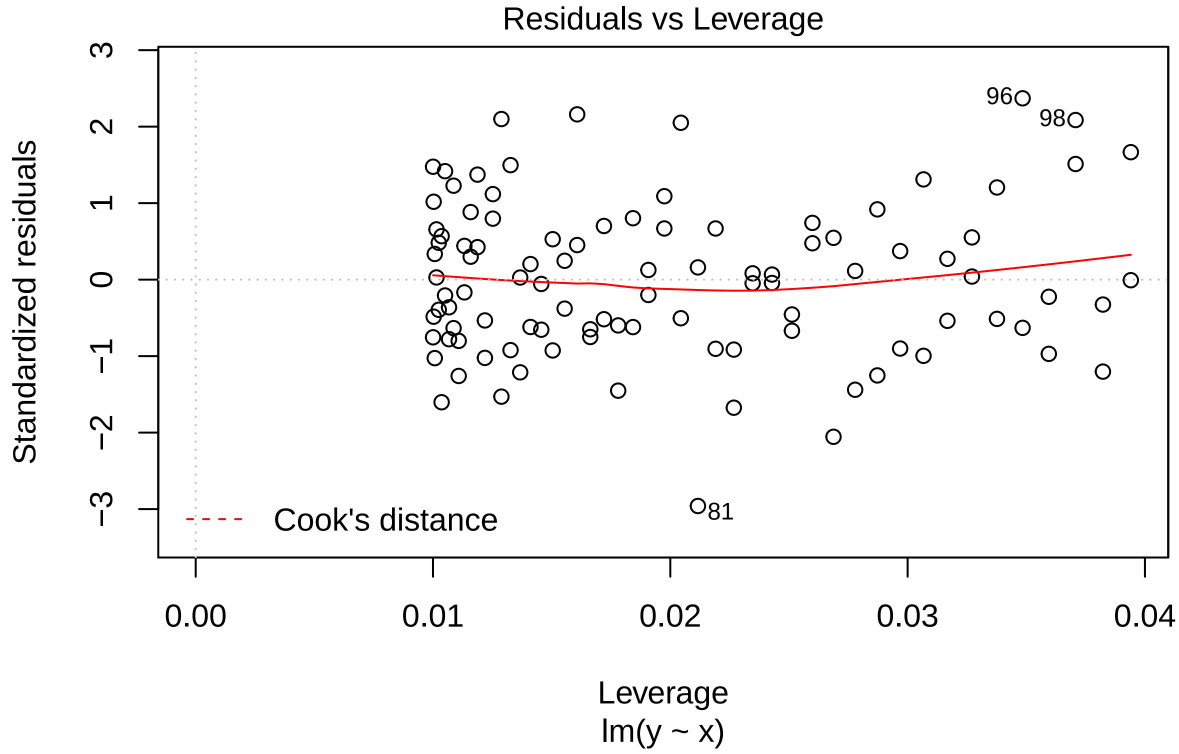
Model Evaluation:

```r
plot(lin)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−30
28
96
81
0
50
100
150
200
Fitted values
lm(y ~ x)

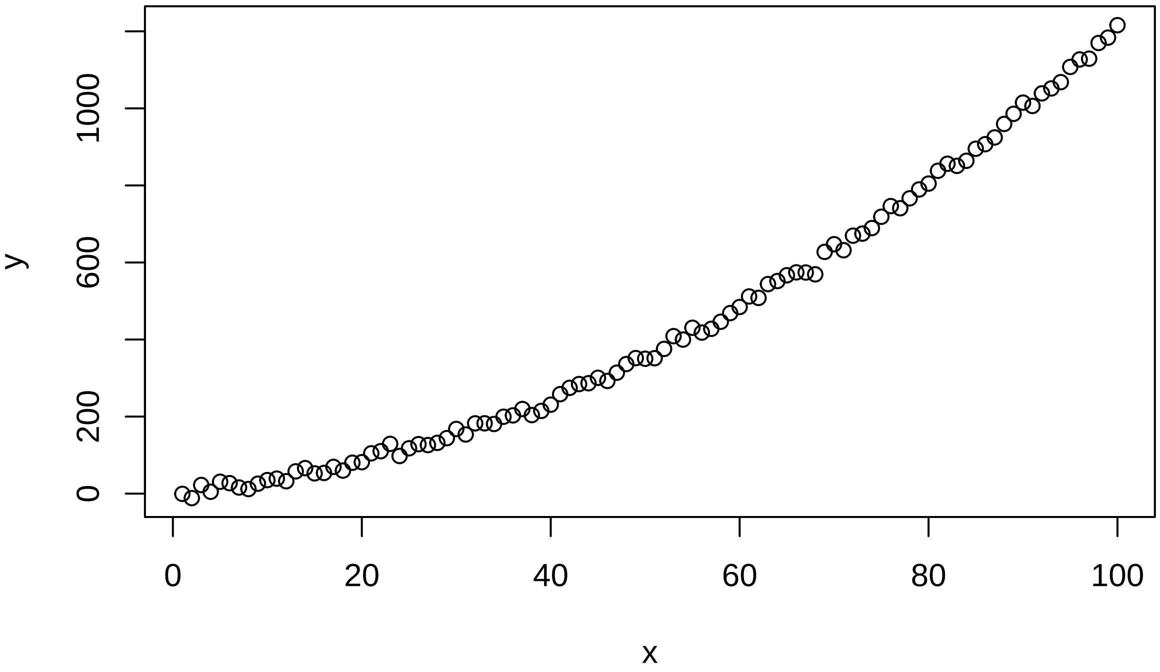

Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
96
28
81
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)





but if model is non-linear:

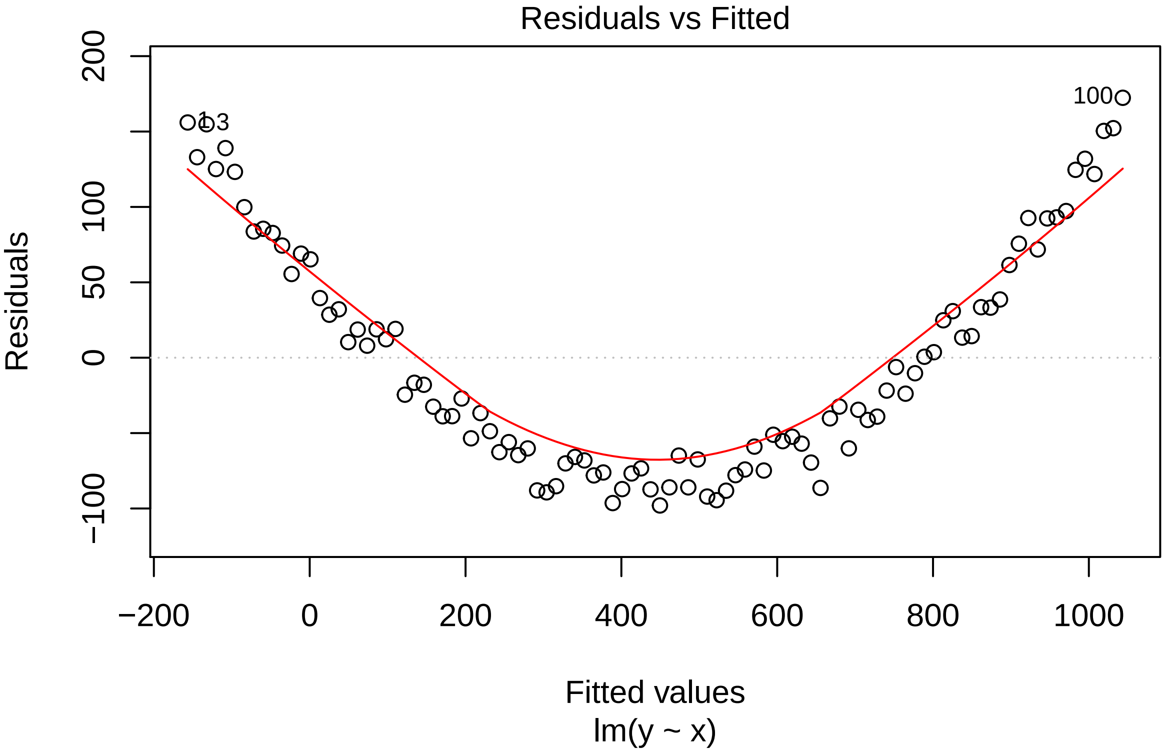
```
x <- seq(1:100)
y <- 4 + 2*x + 0.1*x^2 + 10*rnorm(length(x))
df <- as.data.frame(list(y=y, x=x))
plot(x, y)
```

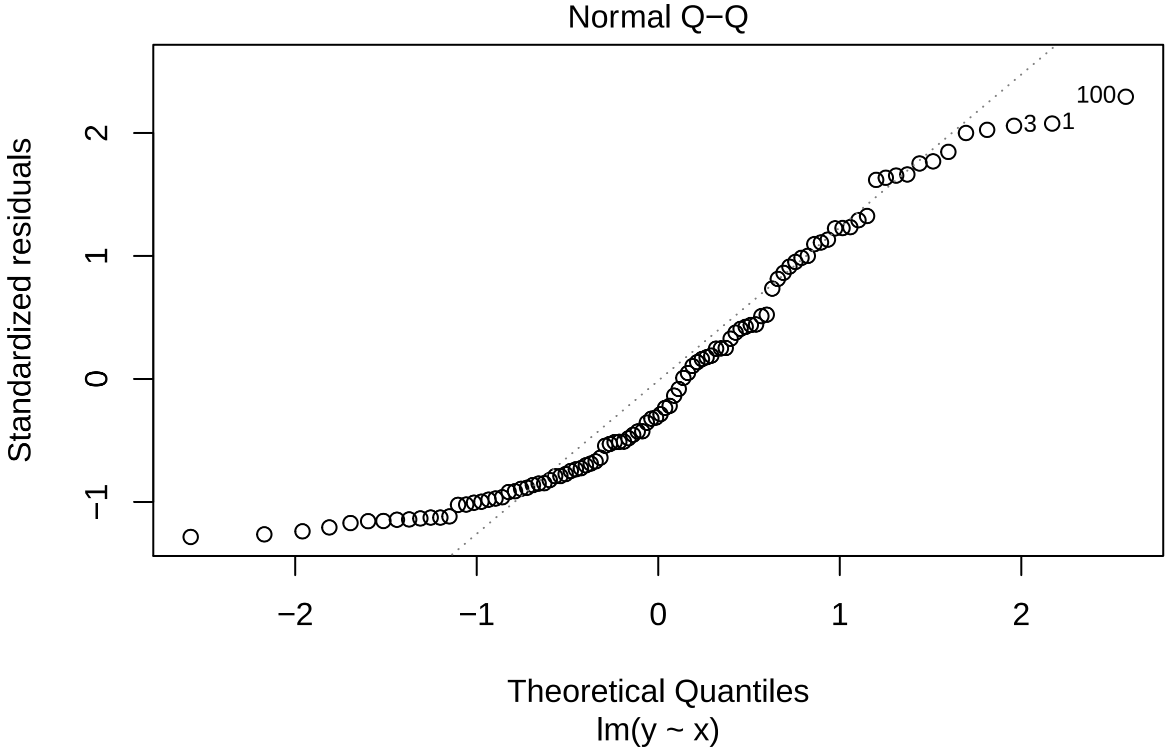


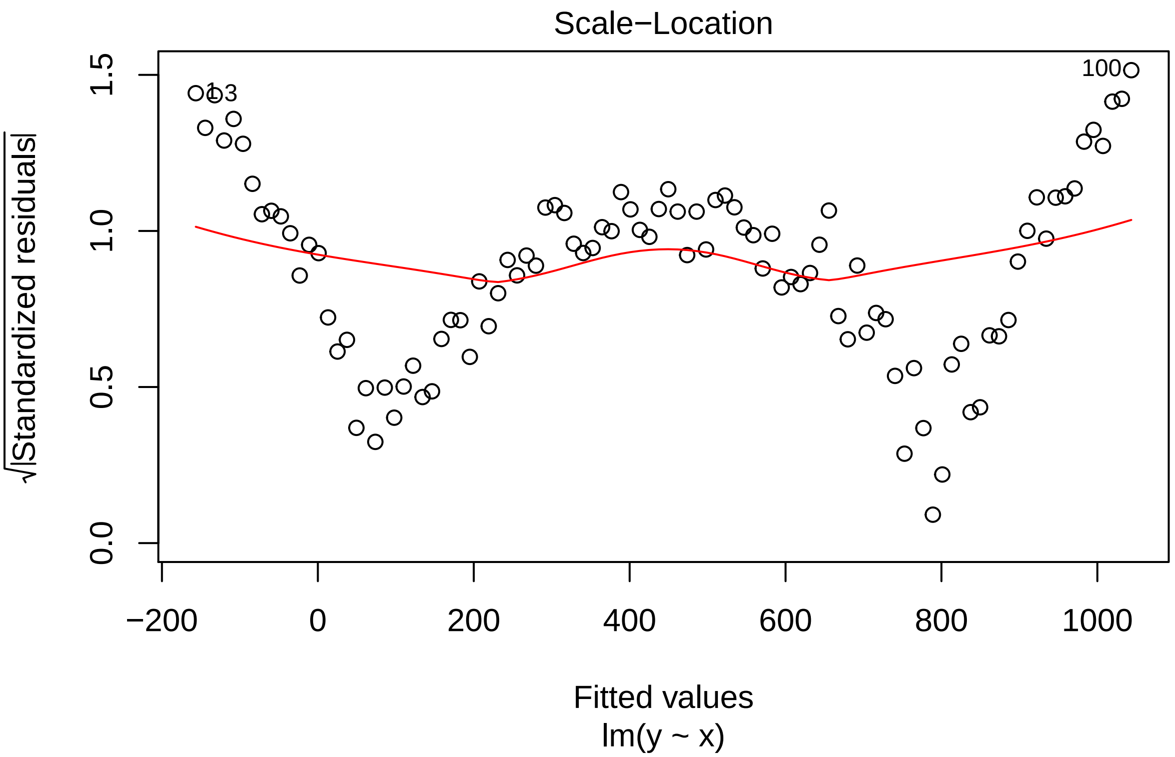
```
lin <- lm(y ~ x, df)
summary(lin)
```

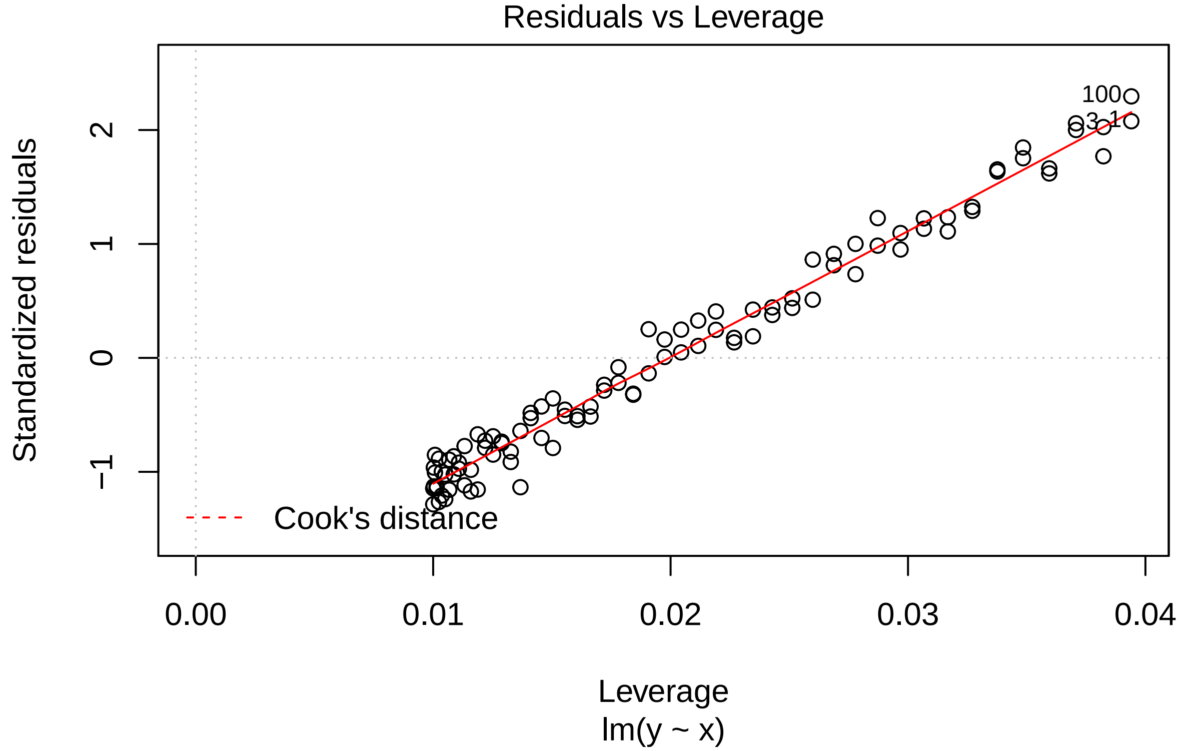
```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -97.99 -65.13 -22.82  62.42 172.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -168.7507    15.4405  -10.93   <2e-16 ***
## x             12.1230     0.2654   45.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.62 on 98 degrees of freedom
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.9547
## F-statistic:  2086 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(lin)
```


Residuals vs Fitted
1
3
100
Residuals
200
100
50
0
−100
−200
0
200
400
600
800
1000
Fitted values
lm(y ~ x)


Normal Q−Q
100
3
1
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)





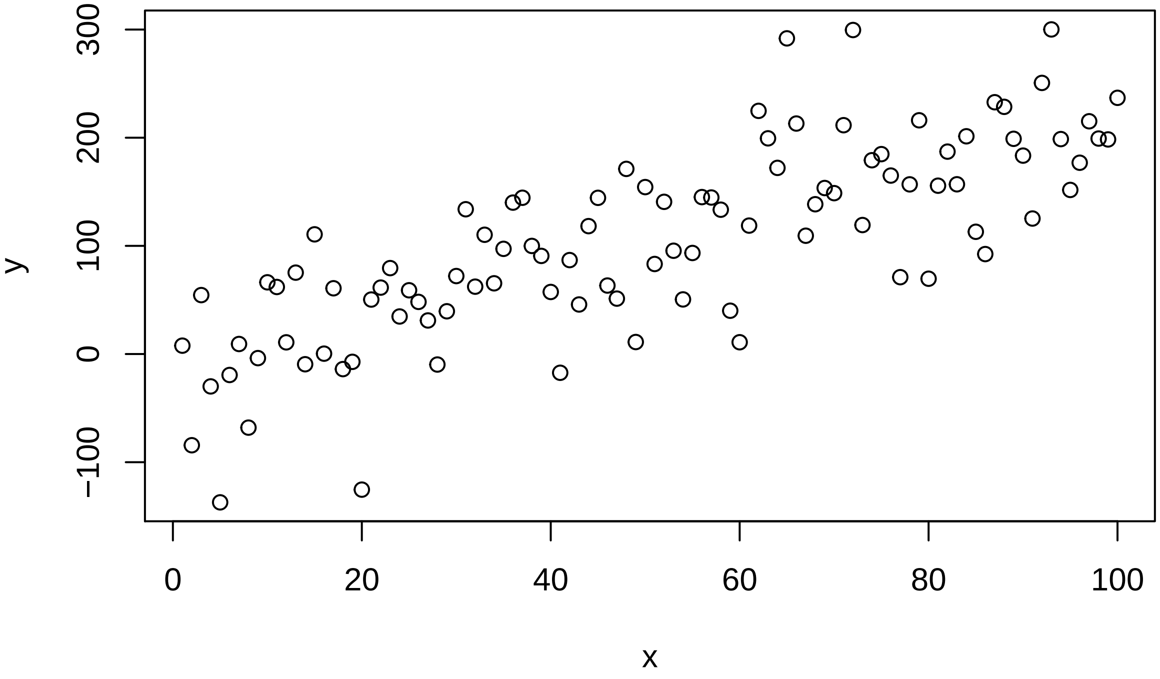
Need to transform variables before model fitting.

R squared

Sometimes there is too much scatter, we can quantify it with coefficient of determination:

$$r^2 = \frac{SSR}{SST}$$

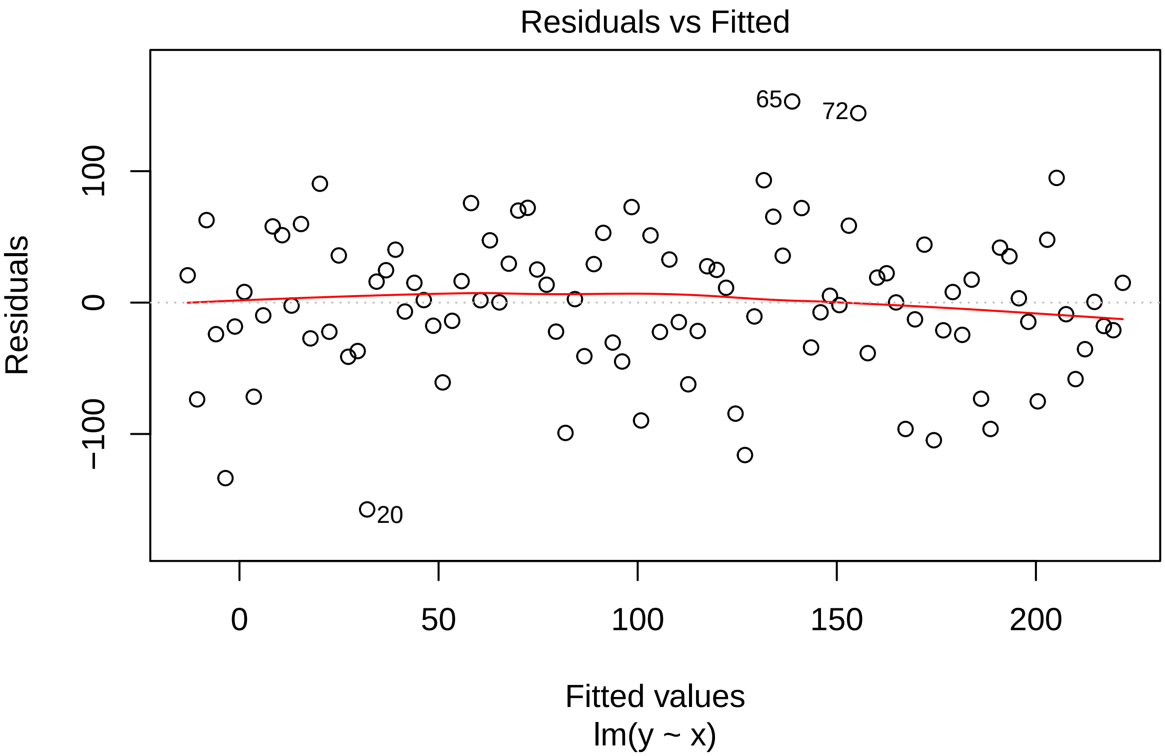
```
x <- seq(1:100)
y <- 4 + 2*x + 50*rnorm(length(x))
plot(x, y)
```

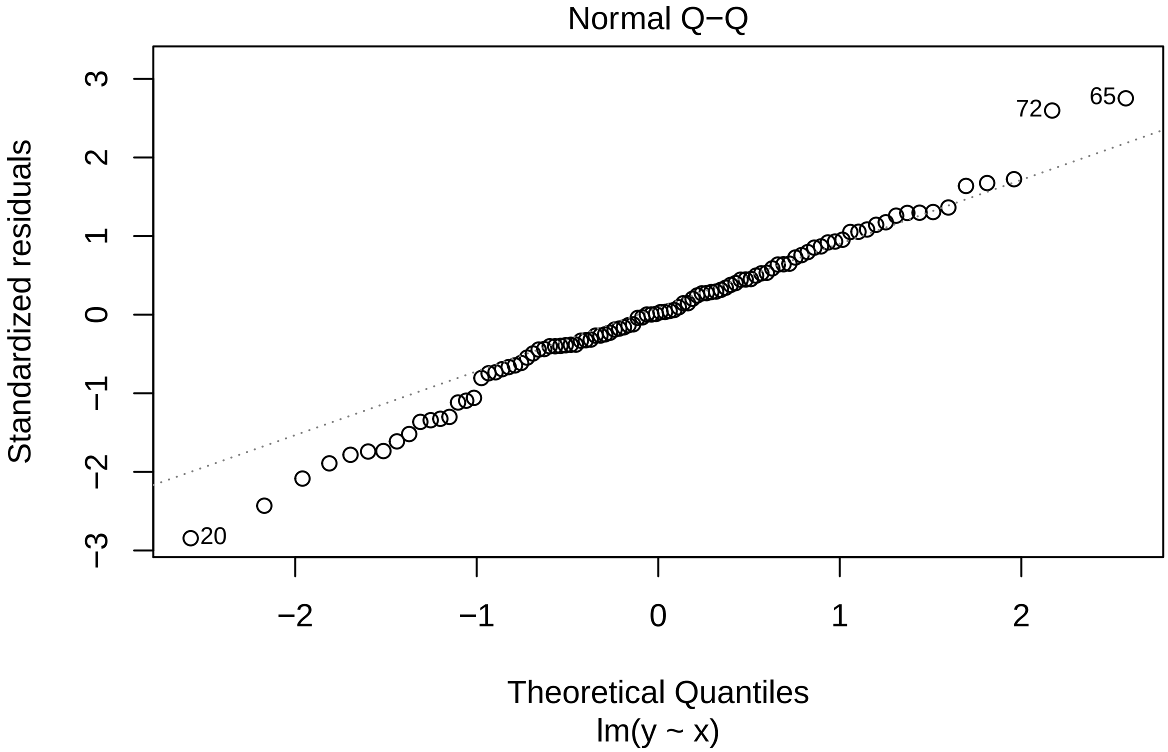


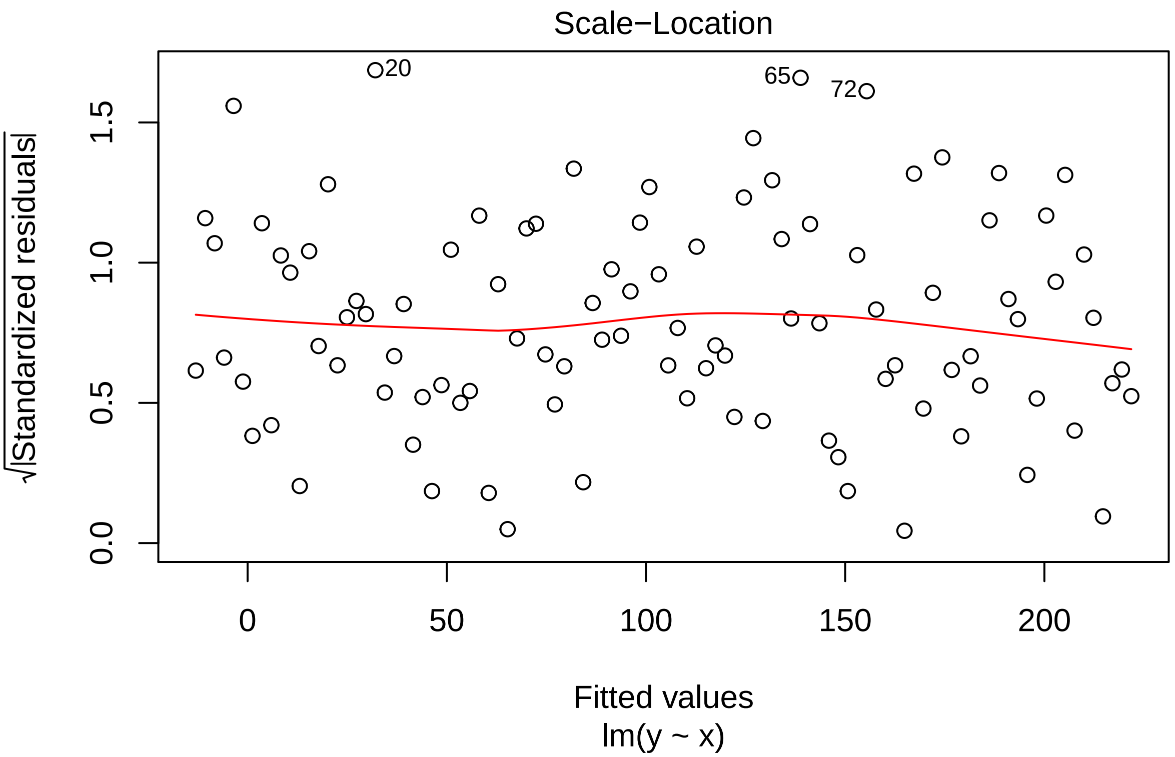
```
df <- as.data.frame(list(y=y, x=x))
lin <- lm(y ~ x, df)
summary(lin)
```

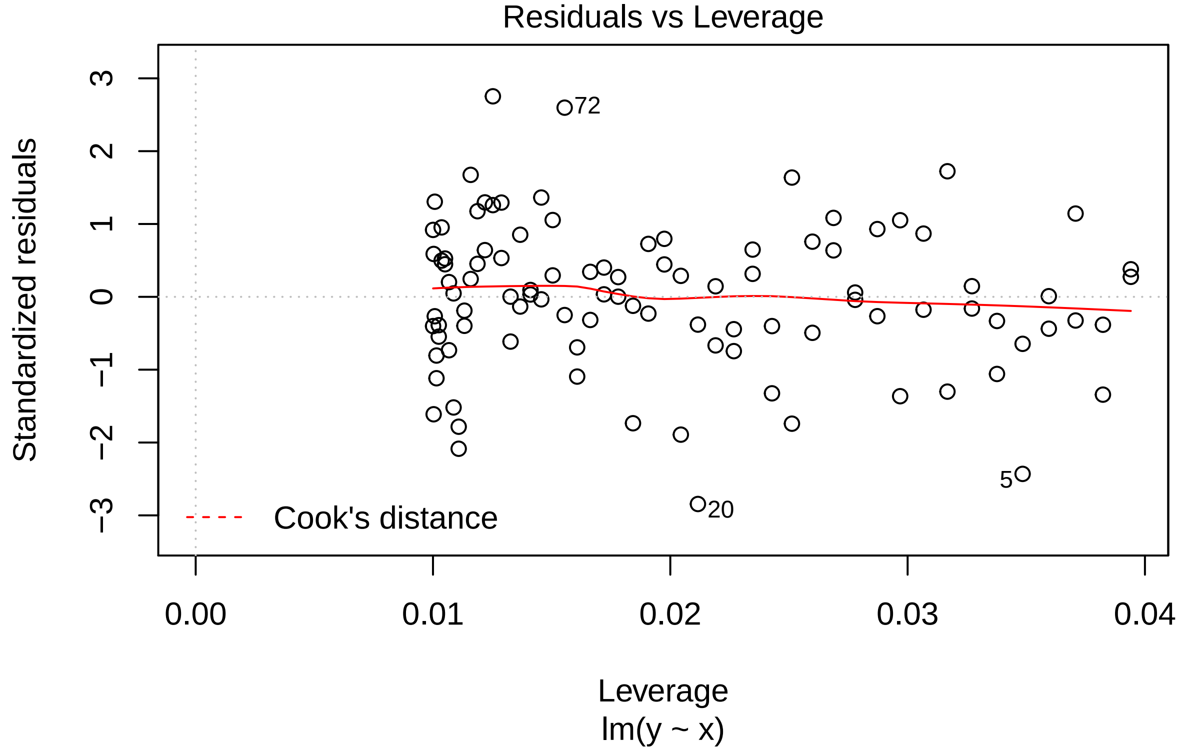
```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -157.410  -25.238    1.138   35.313  153.045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.3773    11.2733  -1.364    0.176
## x             2.3719     0.1938  12.238   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.94 on 98 degrees of freedom
## Multiple R-squared:  0.6045, Adjusted R-squared:  0.6005
## F-statistic: 149.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
plot(lin)
```


Residuals vs Fitted
Residuals
100
0
−100
65
72
20
0
50
100
150
200
Fitted values
lm(y ~ x)


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
72
65
20
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)

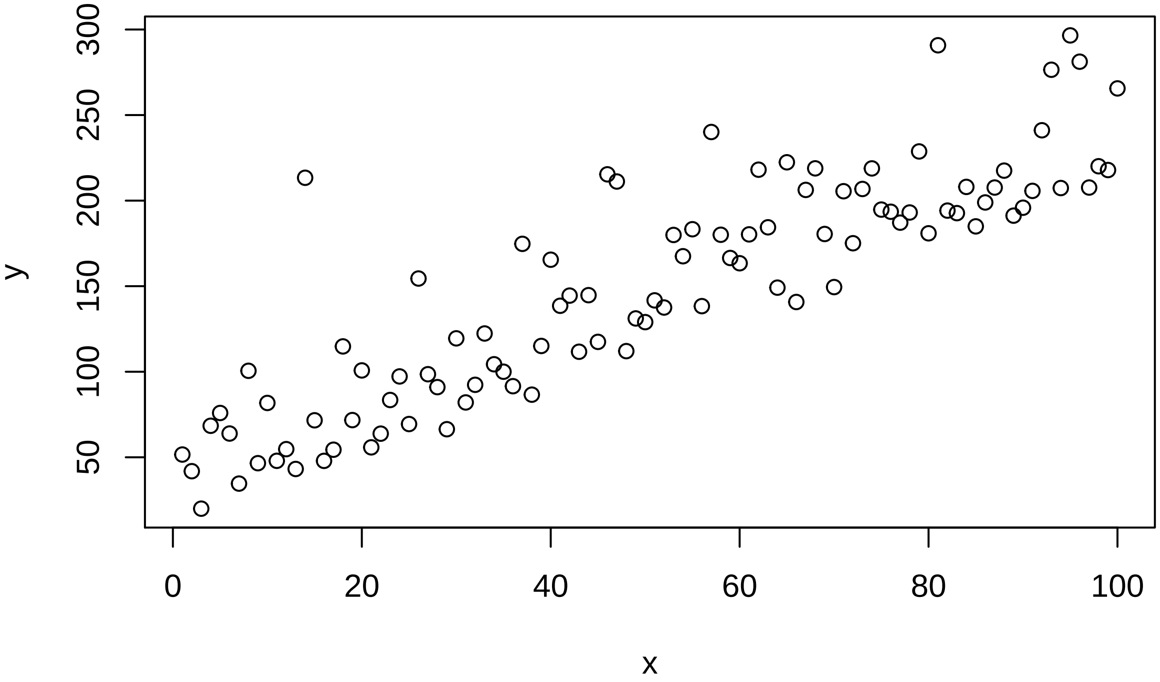




Non-normality

If errors are not normal

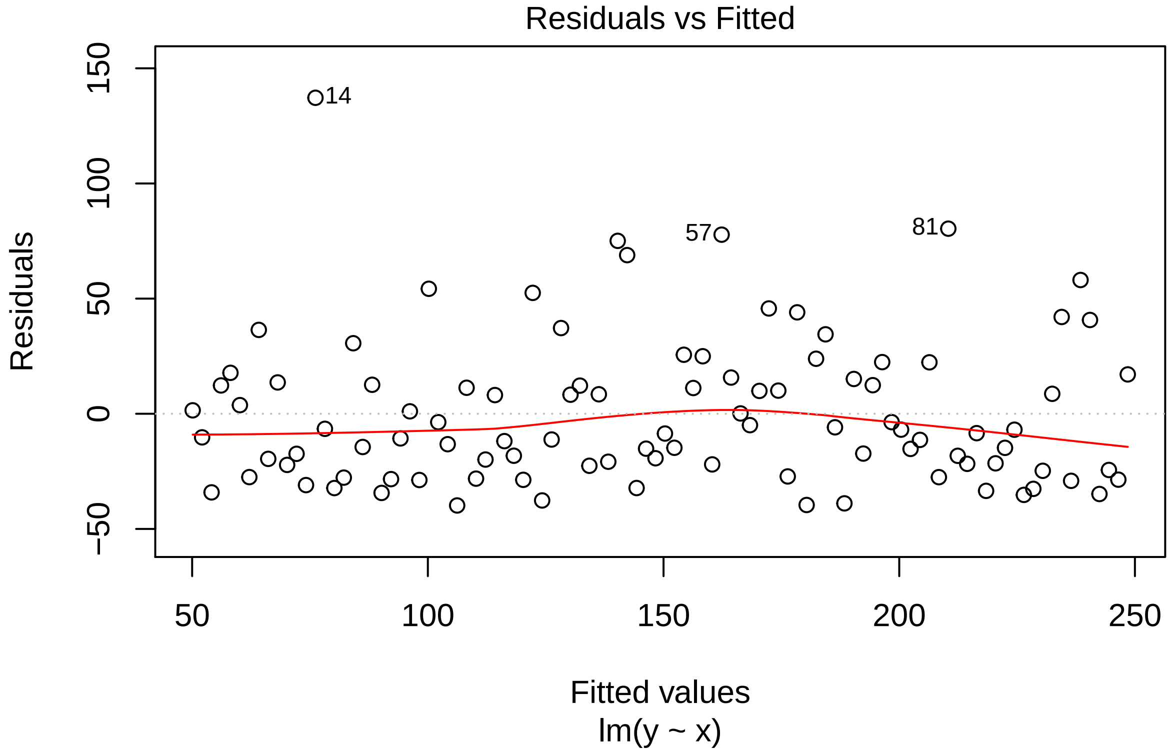
```r
x <- seq(1:100)
y <- 4 + 2*x + 10*rchisq(length(x), df=4)
plot(x, y)
```

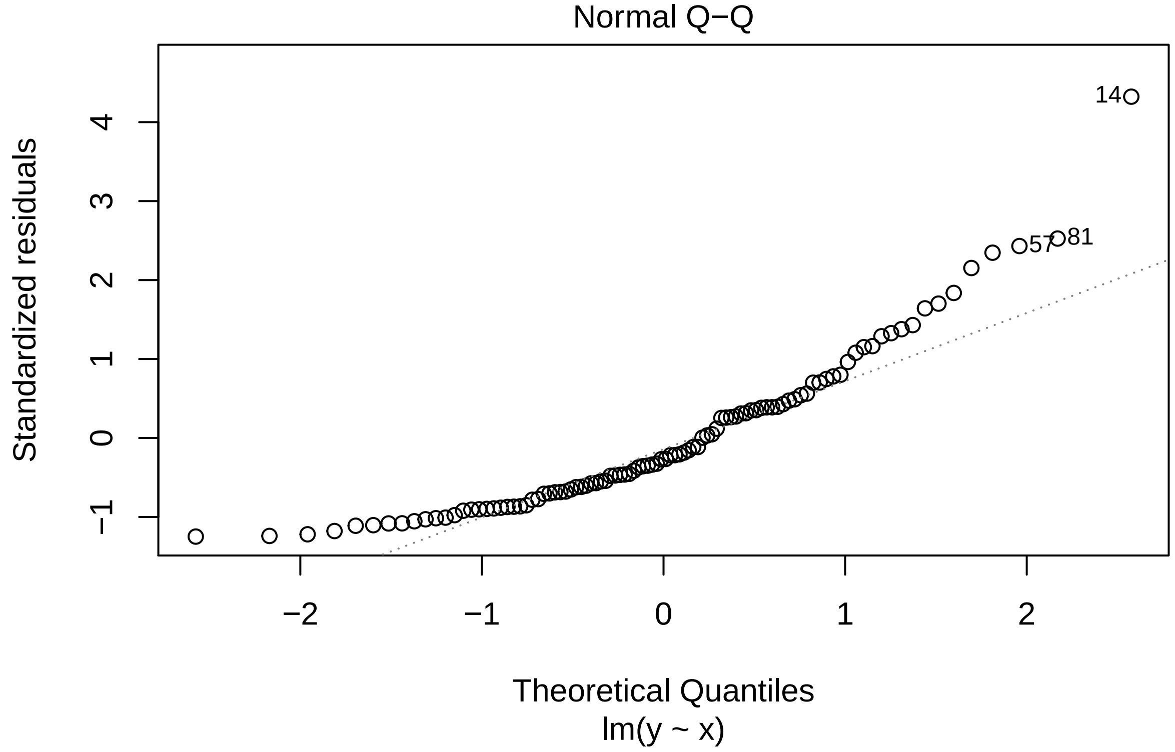


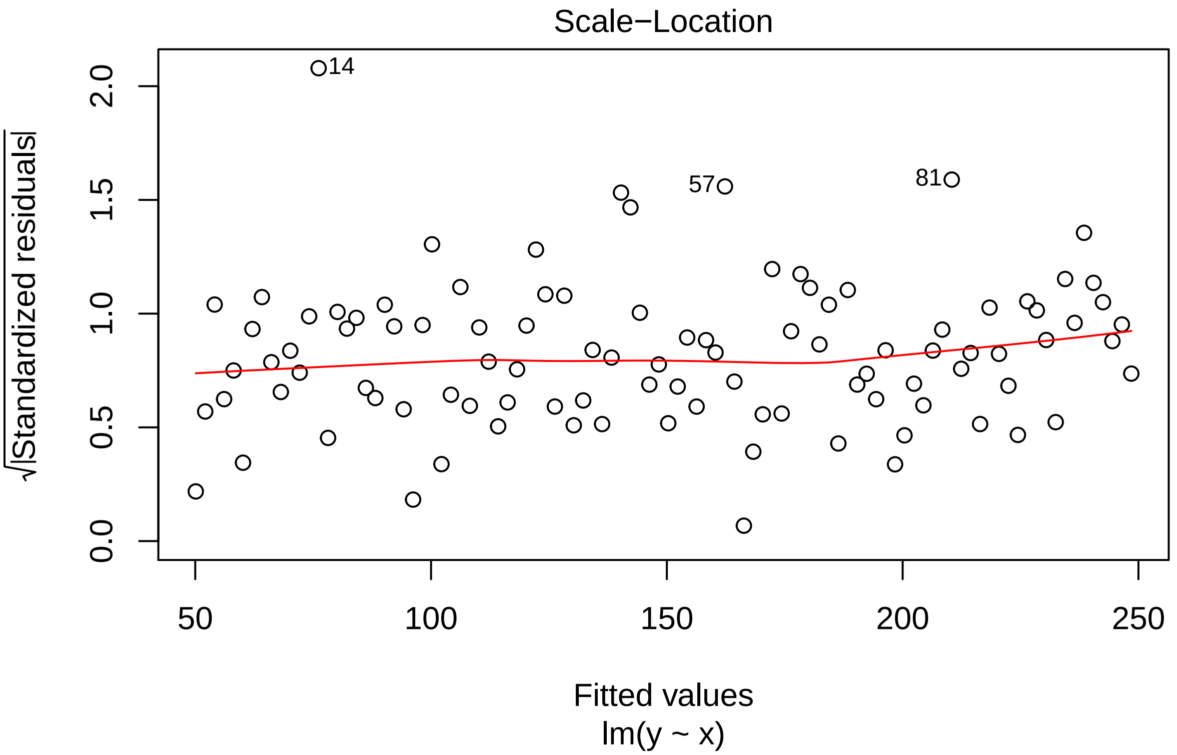
```r
df <- as.data.frame(list(y=y, x=x))
lin <- lm(y ~ x, df)
summary(lin)
```

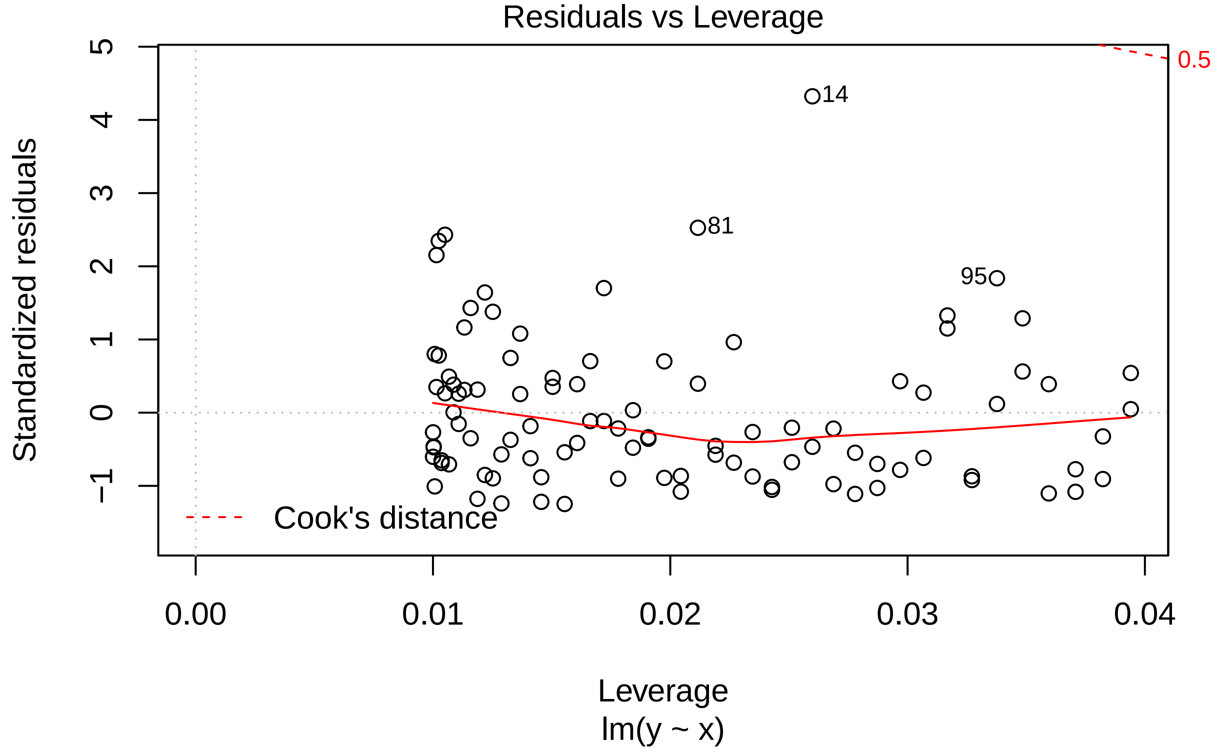
```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39.81 -23.04  -8.50  13.99 137.19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.1066     6.4804   7.423 4.22e-11 ***
## x             2.0038     0.1114  17.986  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.16 on 98 degrees of freedom
## Multiple R-squared:  0.7675, Adjusted R-squared:  0.7651
## F-statistic: 323.5 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
plot(lin)
```


Residuals vs Fitted
Residuals
150
100
50
0
−50
14
57
81
50
100
150
200
250
Fitted values
lm(y ~ x)


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
14
57
81
−2
−1
0
1
2
Theoretical Quantiles
lm(y ~ x)





What if Explanatory Variables are Dependent

If variables are linearly dependent:

```
x1 <- seq(1:100)
x2 <- 10*x1
y <- x1 + x2 + 3 + rnorm(length(x1))
df <- as.data.frame(list(y=y, x1=x1, x2=x2))
lin <- lm(y ~ x1 + x2, df)
summary(lin)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11884 -0.77224 -0.04385  0.69375  3.16182
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.340583   0.210387   15.88   <2e-16 ***
## x1          10.993190   0.003617 3039.40   <2e-16 ***
## x2                 NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.044 on 98 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 9.238e+06 on 1 and 98 DF,  p-value: < 2.2e-16
```

How to find correlated variables:

```
cor(df)
```

```
##            y        x1        x2
## y  1.0000000 0.9999947 0.9999947
## x1 0.9999947 1.0000000 1.0000000
## x2 0.9999947 1.0000000 1.0000000
```