Least Squares Regression

Oliver D'Pug

Data

The **htwt** data set is small, but it is sufficient for demonstrating the different approaches to fitting linear (in the β's) models. The variables in **htwt** are **Height** (inches), **Weight** (pounds), and **Group** (1 = male, 2 = female).

```
htwt <- read.csv("htwt.csv")
htwt$Female <- htwt$Group - 1
summary(htwt)
```

```
##      Height          Weight          Group          Female
##  Min.   :51.0   Min.   : 82.0   Min.   :1.00   Min.   :0.00
##  1st Qu.:56.0   1st Qu.:108.2   1st Qu.:1.00   1st Qu.:0.00
##  Median :59.5   Median :123.5   Median :2.00   Median :1.00
##  Mean   :62.1   Mean   :139.6   Mean   :1.55   Mean   :0.55
##  3rd Qu.:68.0   3rd Qu.:166.8   3rd Qu.:2.00   3rd Qu.:1.00
##  Max.   :79.0   Max.   :228.0   Max.   :2.00   Max.   :1.00
```

Matrix Approach

Using matrix algebra, We can find estimates of the parameters in the theoretical linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} \mathrm{N}\left(0, \sigma^2\right)$.

The data matrix, $\mathbf{X}$, and reponse variable, $\mathbf{Y}$, are defined below.

```
X <- as.matrix(cbind(1, htwt[,c("Height","Female")]))
colnames(X)[1] <- "Intercept"
dim(X)
```

```
## [1] 20  3
```

```
Y <- as.matrix(htwt["Weight"])
dim(Y)
```

```
## [1] 20  1
```

We calculate the **hat** matrix as

```
(hat <- t(X) %*% X)
```

```
##           Intercept Height Female
## Intercept        20   1242     11
## Height         1242  78482    657
## Female           11    657     11
```

Its inverse is

```
  solve(hat)
```

```
##             Intercept        Height       Female
## Intercept  3.58509825 -0.0534459560 -0.392917061
## Height    -0.05344596  0.0008222455  0.004335476
## Female    -0.39291706  0.0043354762  0.224879985
```

Thus, the parameter estimates are

```
  beta <- solve(t(X) %*% X) %*% t(X) %*% Y
  beta
```

```
##                   [,1]
## Intercept -170.699656
## Height       5.010764
## Female      -1.579608
```

lm

R has a number of functions that find parameter estimates for simple linear models. **lm** is one of these functions. We use it here to obtain the estimates — and other useful values as well.

```
  htwt.lm <- lm(Weight ~ Height + factor(Female), data=htwt)
  summary(htwt.lm)
```

```
##
## Call:
## lm(formula = Weight ~ Height + factor(Female), data = htwt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.539 -6.022 -1.253  4.032 14.720
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -170.6997    13.8866 -12.292 6.96e-10 ***
## Height             5.0108     0.2103  23.826 1.68e-14 ***
## factor(Female)1   -1.5796     3.4779  -0.454    0.655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.334 on 17 degrees of freedom
## Multiple R-squared:  0.9741, Adjusted R-squared:  0.9711
## F-statistic: 319.9 on 2 and 17 DF,  p-value: 3.239e-14
```

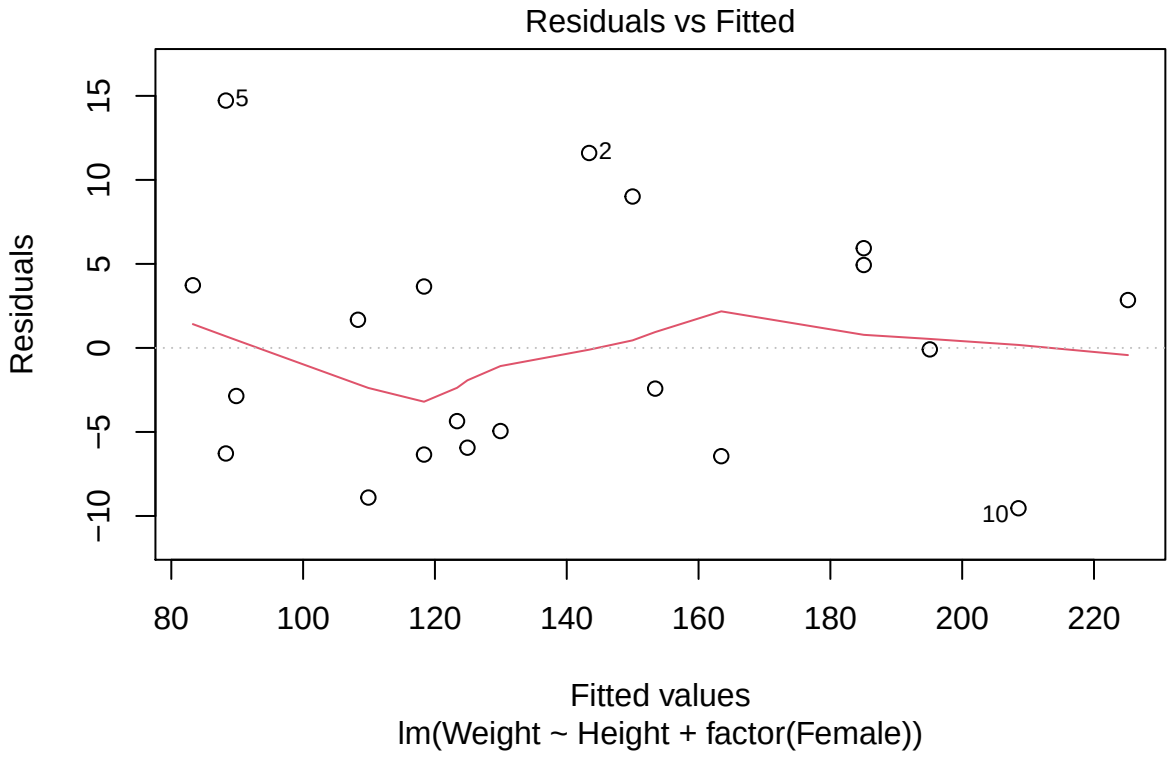
```
  anova(htwt.lm)
```

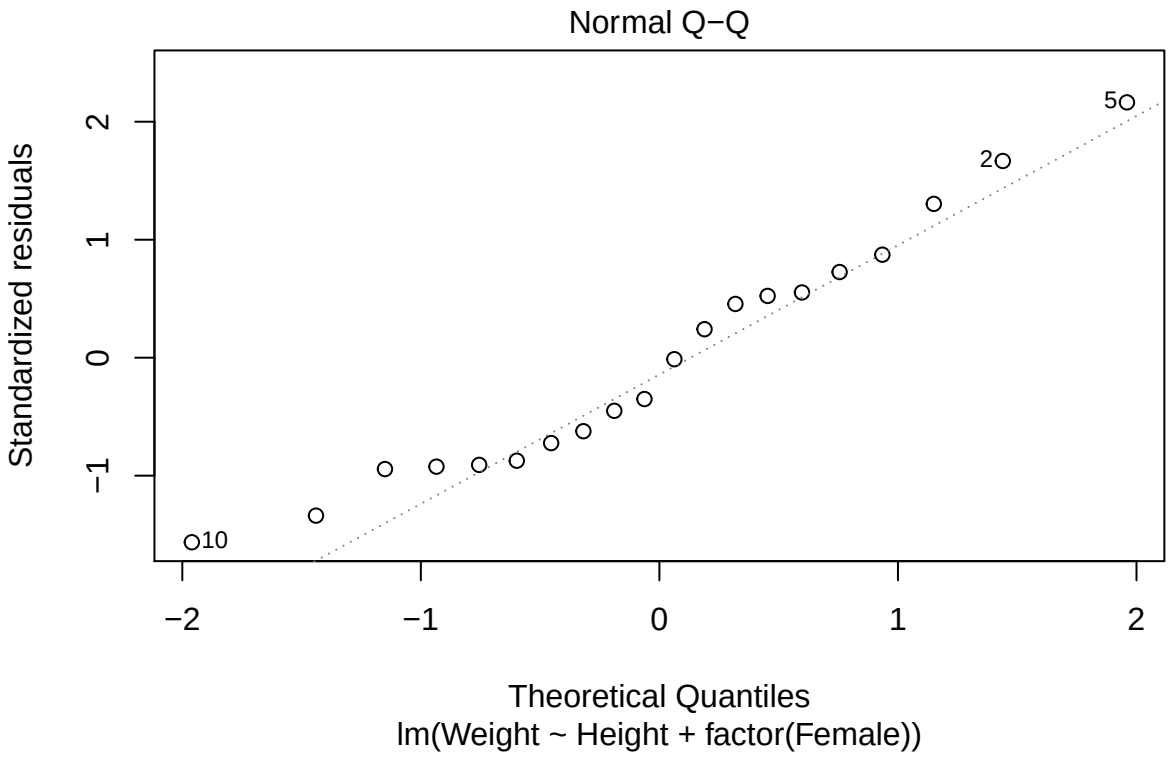
```
## Analysis of Variance Table
##
## Response: Weight
##                Df Sum Sq Mean Sq  F value    Pr(>F)
## Height          1  34405   34405 639.6404 6.265e-15 ***
## factor(Female)  1     11      11   0.2063    0.6554
## Residuals      17    914      54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

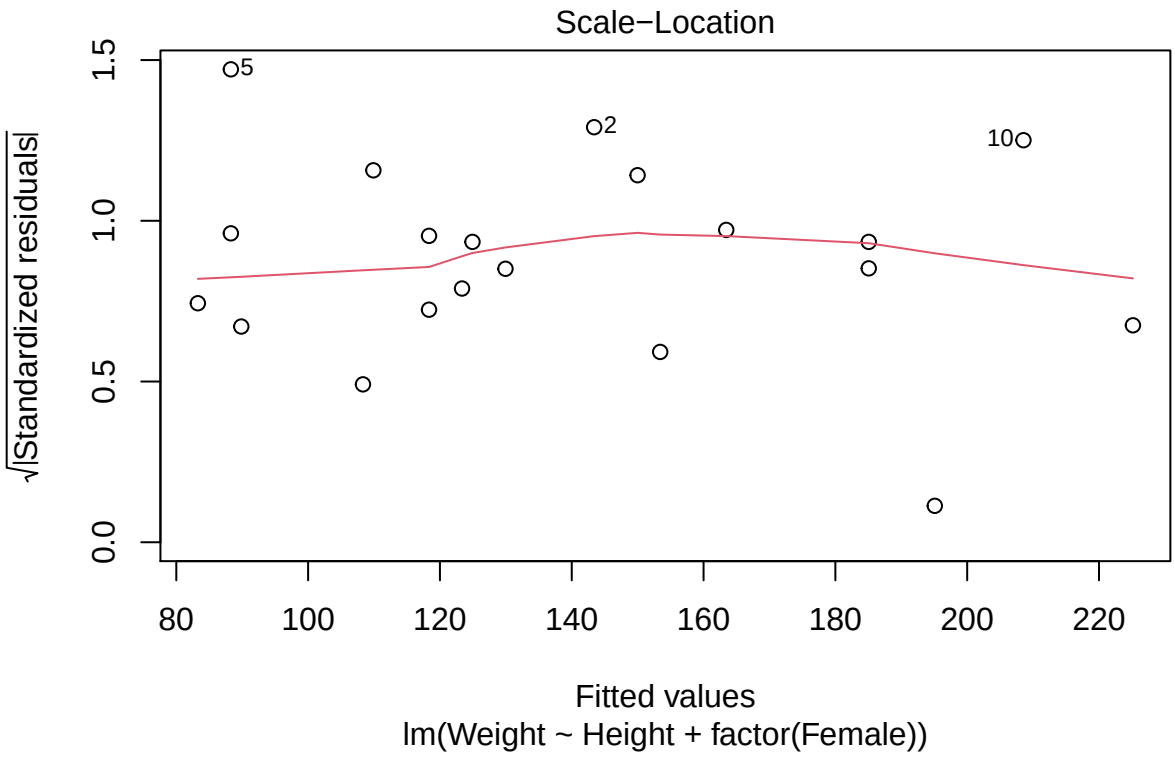
```r
coef(htwt.lm)
```

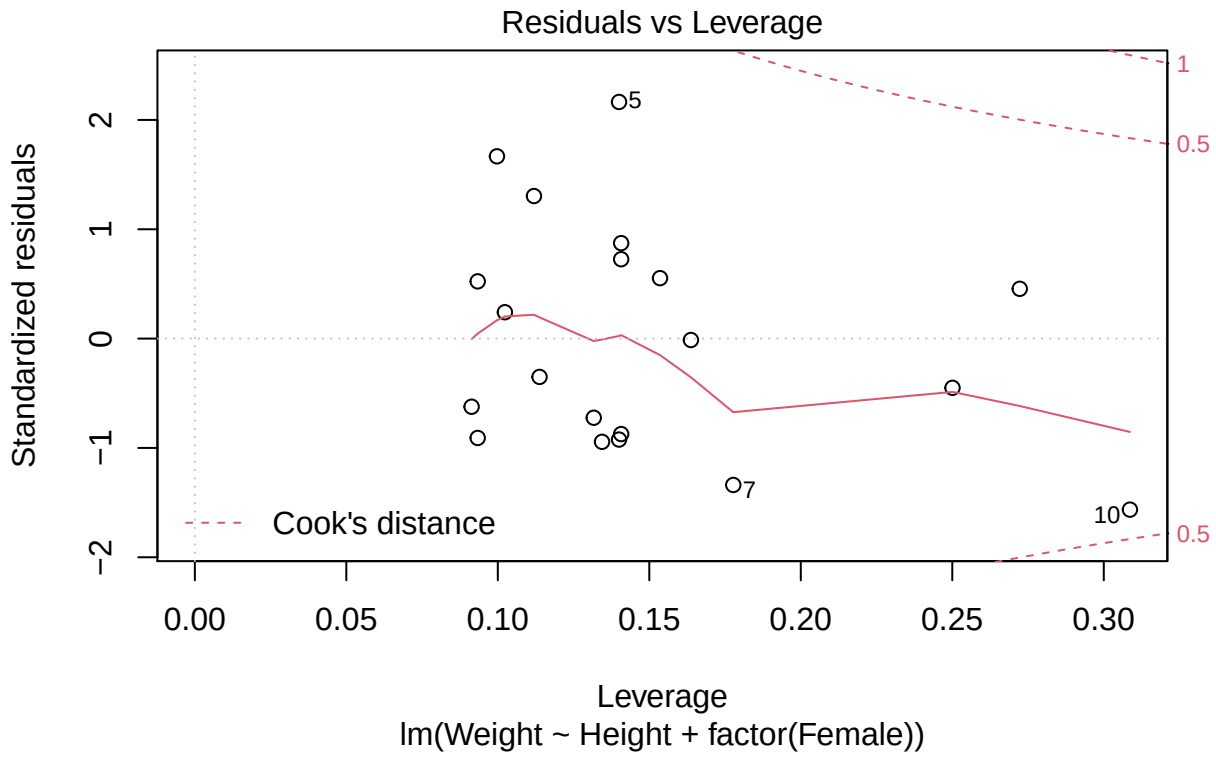
```
##     (Intercept)          Height factor(Female)1
##     -170.699656        5.010764       -1.579608
```

```r
plot(htwt.lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Weight ~ Height + factor(Female))
10
2
5
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
5
2
10
80
100
120
140
160
180
200
220
Fitted values
lm(Weight ~ Height + factor(Female))



Interactions

It may be that males and females add weight at different rates relative to height. We can examine this by including an interaction effect.

```
  head(X)
```

```
##      Intercept Height Female
## [1,]         1     64      0
## [2,]         1     63      1
## [3,]         1     67      1
## [4,]         1     60      0
## [5,]         1     52      1
## [6,]         1     58      1
```

```
  X2 <- cbind(X, X[,"Height"] * X[,"Female"])
  colnames(X2)[4] <- "HtFem"
  head(X2)
```

```
##      Intercept Height Female HtFem
## [1,]         1     64      0     0
## [2,]         1     63      1    63
## [3,]         1     67      1    67
## [4,]         1     60      0     0
## [5,]         1     52      1    52
## [6,]         1     58      1    58
```

```
  (hat <- t(X2) %*% X2)
```

```
##           Intercept Height Female HtFem
## Intercept        20   1242     11   657
## Height         1242  78482    657 39813
## Female           11    657     11   657
## HtFem           657  39813    657 39813
```

```
  solve(t(X2) %*% X2) %*% t(X2) %*% Y
```

```
##                    [,1]
## Intercept -198.2608696
## Height       5.4347826
## Female      54.4858457
## HtFem       -0.9012586
```

```
  htwt.lm2 <- lm(Weight ~ Height + factor(Female) + Height:factor(Female), data=htwt)
  summary(htwt.lm2)
```

```
##
## Call:
## lm(formula = Weight ~ Height + factor(Female) + Height:factor(Female),
##     data = htwt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.968 -3.413 -1.104  2.697 13.163
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -198.2609    16.6933 -11.877 2.39e-09 ***
## Height                    5.4348     0.2547  21.340 3.51e-13 ***
## factor(Female)1          54.4858    23.2997   2.338   0.0327 *
## Height:factor(Female)1   -0.9013     0.3713  -2.427   0.0274 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.463 on 16 degrees of freedom
## Multiple R-squared:  0.9811, Adjusted R-squared:  0.9775
## F-statistic: 276.6 on 3 and 16 DF,  p-value: 5.425e-14
```

```
  anova(htwt.lm2)
```

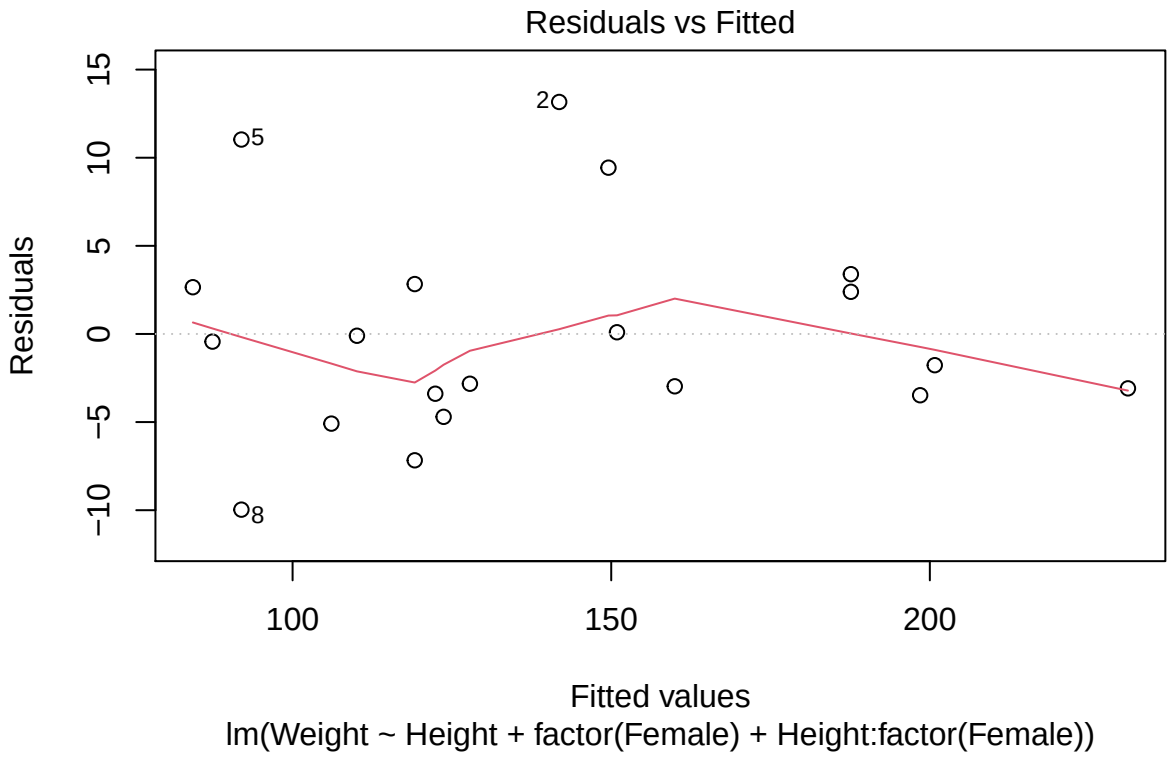
```
## Analysis of Variance Table
##
## Response: Weight
##                        Df Sum Sq Mean Sq  F value    Pr(>F)
## Height                  1  34405   34405 823.7096 3.441e-15 ***
## factor(Female)          1     11      11   0.2656   0.61332
## Height:factor(Female)   1    246     246   5.8921   0.02738 *
## Residuals              16    668      42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

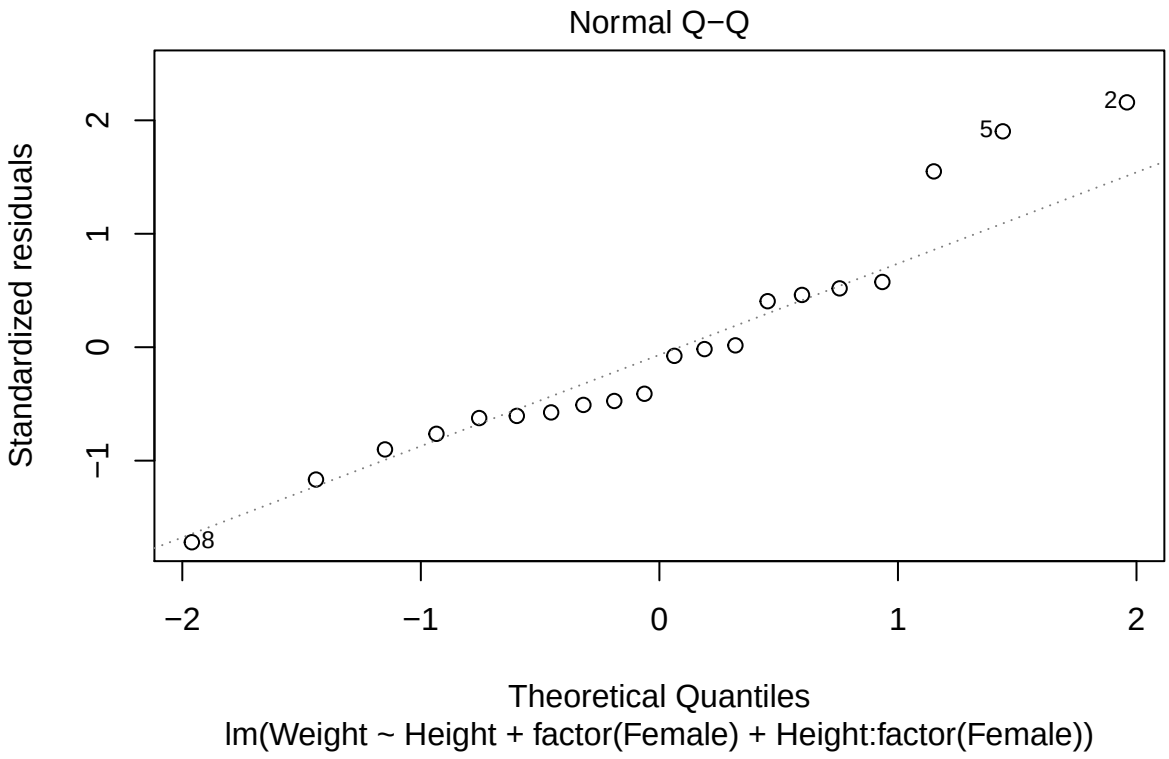
```
  coef(htwt.lm2)
```

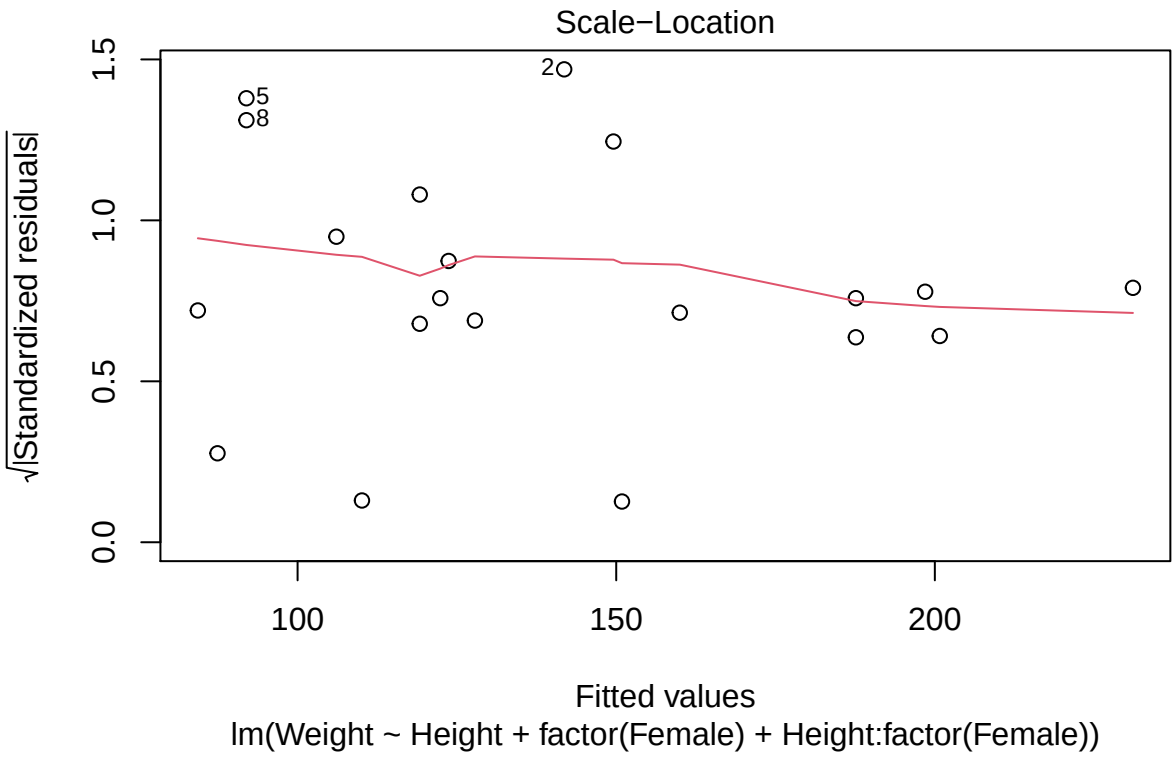
```
##            (Intercept)                 Height        factor(Female)1
```

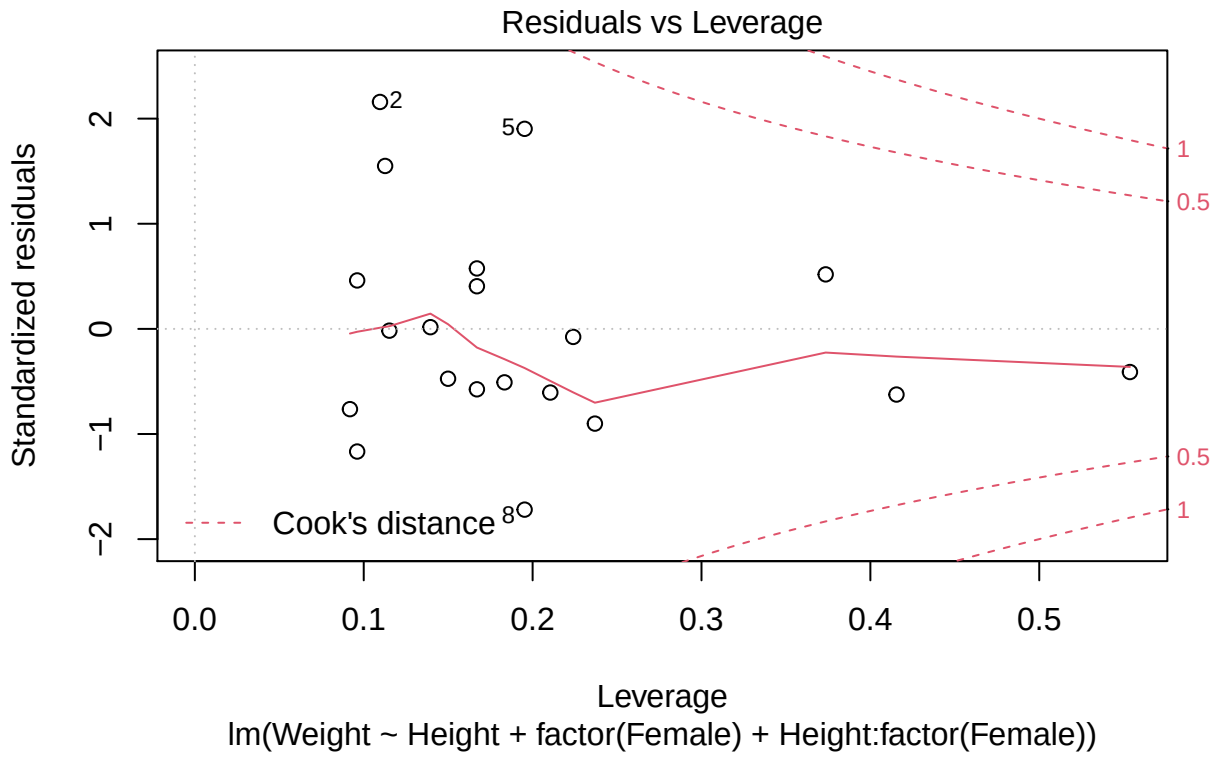
```
##           -198.2608696              5.4347826             54.4858457
## Height:factor(Female)1
##             -0.9012586
```

```
plot(htwt.lm2)
```



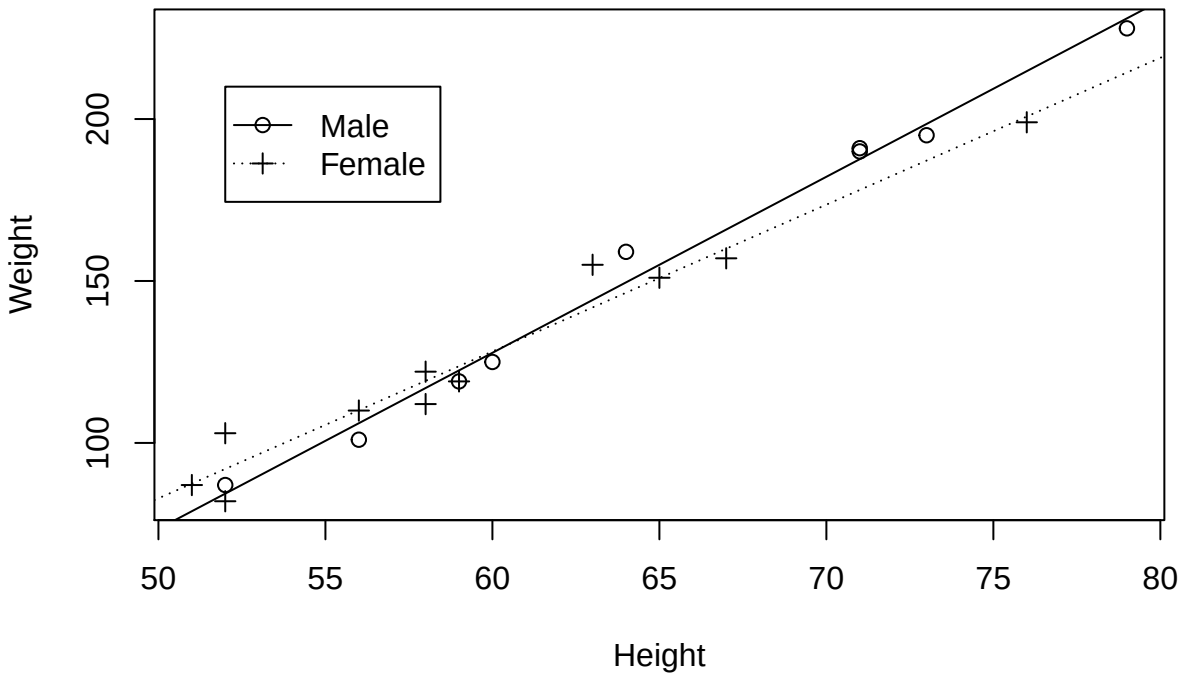

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Weight ~ Height + factor(Female) + Height:factor(Female))
2
5
8
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2
5
8
100
150
200
Fitted values
lm(Weight ~ Height + factor(Female) + Height:factor(Female))



A plot of the data shows that the interaction exists.

```
with(htwt,
     {
       plot(Height, Weight, type="n")
       points(Height[Group==1], Weight[Group==1], pch=1)
       points(Height[Group==2], Weight[Group==2], pch=3)
       abline(reg=lm(Weight[Group==1] ~ Height[Group==1]), lty=1)
       abline(reg=lm(Weight[Group==2] ~ Height[Group==2]), lty=3)
       legend(52, 210, legend=c("Male","Female"), pch=c(1,3), lty=c(1,3))
     }
)
```


Male
Female
Weight
Height
100
150
200
50
55
60
65
70
75
80