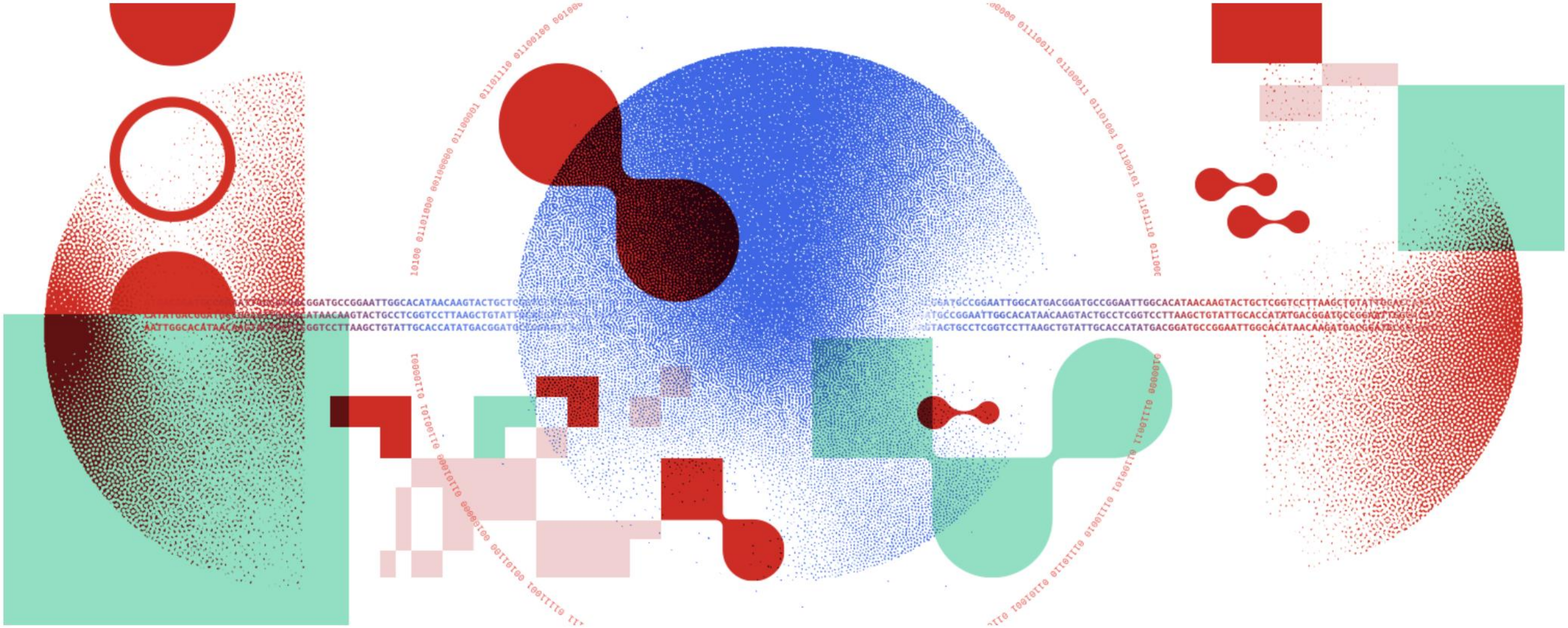


Introduction to Statistics and Data Visualisation with R

Lausanne, January 2026

Rachel Marcone and Thomas Zwahlen

Multiple Regression

What happens if both,
age and weight variables
were included in the same model ?

One multiple regression with two variables

```
Call:
lm(formula = Height ~ Age + Weight, data = class)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6248 -1.3016 -0.0176  0.8324  4.1019 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 132.1943     5.0823  26.011 1.61e-14 ***
Age           1.2267     0.5302   2.314  0.03431 *  
Weight        0.2761     0.0695   3.973  0.00109 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.255 on 16 degrees of freedom
Multiple R-squared:  0.828,	Adjusted R-squared:  0.8065 
F-statistic: 38.52 on 2 and 16 DF,  p-value: 7.646e-07
```

This model allows us to determine the respective contribution of each variable separately.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 132.1943     5.0823  26.011 1.61e-14 ***
Age           1.2267     0.5302   2.314  0.03431 *  
Weight        0.2761     0.0695   3.973  0.00109 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is similar to the simple regression case.

Each test is conducted assuming that the tested parameter is the last one entering the model:

« If *weight* is already in the model, is the coefficient for *age* significantly different from 0 ? »

Two single regressions vs one multiple regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 142.57014    2.67989  53.200  < 2e-16 ***
Weight        0.39523    0.05231   7.555 7.89e-07 ***
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 125.2239     6.5217  19.201 5.82e-13 ***
Age           2.7871     0.4869   5.724 2.48e-05 ***
---
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 132.1943     5.0823  26.011 1.61e-14 ***
Age           1.2267     0.5302   2.314  0.03431 *
Weight        0.2761     0.0695   3.973  0.00109 **
---
```

While both age and weight seem significant by themselves, age is much less significant when weight is already included (see also the R^2).

It is likely that a lot of the information provided by the age is also provided by the weight, so that there may be little need to have both terms in the model.

```
Multiple R-squared: 0.828,      Adjusted R-squared: 0.8065
```

As before, R^2 is the proportion of the total variance in the response data that is explained by the model.

Adding a new variable in the model will always increase R^2, up to 1 when there the number of degrees of freedom is 0 (number of parameters to estimate = number of observations).

Multiple and adjusted R-squared

```
Multiple R-squared: 0.828,      Adjusted R-squared: 0.8065
```

The adjusted R-squared adjusts for the number of variables in the model, and does not necessarily increase when the number of variables increase; it can even be negative.

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

It is always equal or below R².

$$\text{Adjusted R}^2 = 1 - \frac{SS_{residuals} / (n-K)}{SS_{total} / (n-1)}$$

Example

```
y <- rnorm(10)
x1 <- rnorm(10); x2 <- rnorm(10); … ; x9 <-
rnorm(10)
summary(lm(y ~ x1)); summary(lm(y ~ x1+x2));
```

```
1: Multiple R-squared: 0.1419,     Adjusted R-squared: 0.03464
2: Multiple R-squared: 0.5173,     Adjusted R-squared: 0.3794
3: Multiple R-squared: 0.557,      Adjusted R-squared: 0.3355
4: Multiple R-squared: 0.5577,     Adjusted R-squared: 0.2039
5: Multiple R-squared: 0.7953,     Adjusted R-squared: 0.5395
6: Multiple R-squared: 0.8321,     Adjusted R-squared: 0.4962
7: Multiple R-squared: 0.984,      Adjusted R-squared: 0.9281
8: Multiple R-squared: 0.9851,     Adjusted R-squared: 0.866
9: Multiple R-squared:     1,      Adjusted R-squared:   NaN
```

The last regression from the example

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)

Residuals:
ALL 10 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02693         NA      NA       NA
x1           0.53886         NA      NA       NA
x2          -0.52227         NA      NA       NA
x3           0.51881         NA      NA       NA
x4           0.74757         NA      NA       NA
x5           0.14394         NA      NA       NA
x6          -0.65387         NA      NA       NA
x7          -0.48271         NA      NA       NA
x8          -0.62487         NA      NA       NA
x9           0.23759         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:     1,      Adjusted R-squared:   NaN 
F-statistic:   NaN on 9 and 0 DF,  p-value: NA
```

F-statistic for significance of regression

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.77355   12.90896   6.335 9.92e-06 ***
Age          3.11575    1.34668   2.314  0.03431 *
Weight       0.35064    0.08827   3.973  0.00109 **

F-statistic: 38.52 on 2 and 16 DF,  p-value: 7.646e-07
```

Again, the F-statistic allows us to test if the whole regression (adding all variables *vs* having only the intercept in) is significant.

If any of the tests for the individual variables is significant, the F-test will generally be significant as well.

However, even if no individual variable is significant (e.g. $p < 0.05$), the F-test can still be significant.

Categorical variables, dummy variables and contrasts

We'd like to use categorical variables in a linear model, as in:

$$\textbf{Height} = b_0 + b_1 \textbf{ Age} + b_2 \textbf{ « Gender » + error}$$

Intuitively, we want to estimate a « Male » and a « Female » effect.

We'd like to use categorical variables in a linear model, as in:

$$\textbf{Height} = b_0 + b_1\,\textbf{Age} + b_2\,\text{« }\textbf{Gender}\text{ »} + \textbf{error}$$

Intuitively, we want to estimate a « Male » and a « Female » effect.

In practice, categorical variables (factors in R) are turned (by default, based on alphabetical order) into **dummy variables** of the form

$$\text{Gender} = \begin{cases} 1 \text{ if } \textbf{F}\text{emale} \\ 2 \text{ if } \textbf{M}\text{ale} \end{cases}$$

Example of summary results of the `lm` command in R

```
Call:
lm(formula = Height ~ Age + Gender, data = class)

Residuals:
   Min     1Q Median     3Q    Max
-3.483 -1.910 -0.319  1.326  5.317

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.5241     5.8886  21.147 4.04e-13 ***
Age           2.7276     0.4398   6.202 1.27e-05 ***
GenderM       2.8362     1.2797   2.216   0.0415 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.78 on 16 degrees of freedom
Multiple R-squared:  0.7387, Adjusted R-squared:  0.706
F-statistic: 22.61 on 2 and 16 DF,  p-value: 2.176e-05
```

Example of summary results of the `lm` command in R

```
Call:
lm(formula = Height ~ Age + Gender, data = class)

Residuals:
   Min     1Q Median     3Q    Max
-3.483 -1.910 -0.319  1.326  5.317

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.5241     5.8886  21.147 4.04e-13 ***
Age           2.7276     0.4398   6.202 1.27e-05 ***
GenderM       2.8362     1.2797   2.216   0.0415 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.78 on 16 degrees of freedom
Multiple R-squared:  0.7387, Adjusted R-squared:  0.706
F-statistic: 22.61 on 2 and 16 DF,  p-value: 2.176e-05
```

baseline for height among Female

Example of summary results of the `lm` command in R

```
Call:
lm(formula = Height ~ Age + Gender, data = class)

Residuals:
   Min     1Q Median     3Q    Max
-3.483 -1.910 -0.319  1.326  5.317

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 124.5241     5.8886  21.147 4.04e-13 ***
Age           2.7276     0.4398   6.202 1.27e-05 ***
GenderM       2.8362     1.2797   2.216   0.0415 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.78 on 16 degrees of freedom
Multiple R-squared:  0.7387, Adjusted R-squared:  0.706
F-statistic: 22.61 on 2 and 16 DF,  p-value: 2.176e-05
```

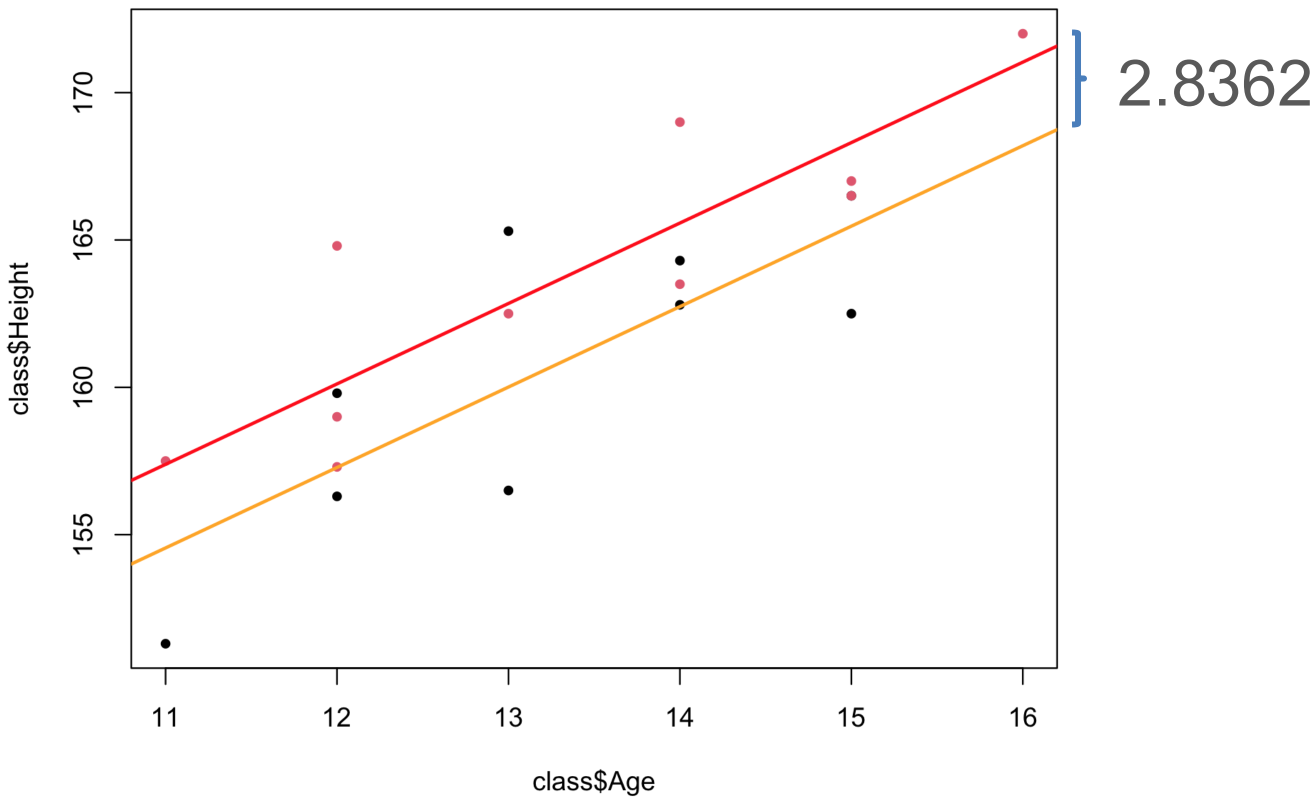
baseline for height among Female

The factor GenderM corresponds to the difference in baseline for Males compared to females

The model specifies 2 straight lines, with the same slope but different y-intercepts:

For women: Height = 124.52 + 2.72 Age (in orange)

For men: Height = 127.3 + 2.72 Age (in red)



We could also compute the difference in means between males and females directly:

```
> tapply(class$Height,class$Gender,mean)
       F        M
160.5889 163.9100
> means <- tapply(class$Height,class$Gender,mean)
> diff(means)
       M
3.321111
```

This result is slightly different from the 2.8362 cm difference found with the linear model.

Where does the difference come from ?

So far, we have assumed a difference between the lines, but the same slope; that is, for both men and women, the effect of age is the same.

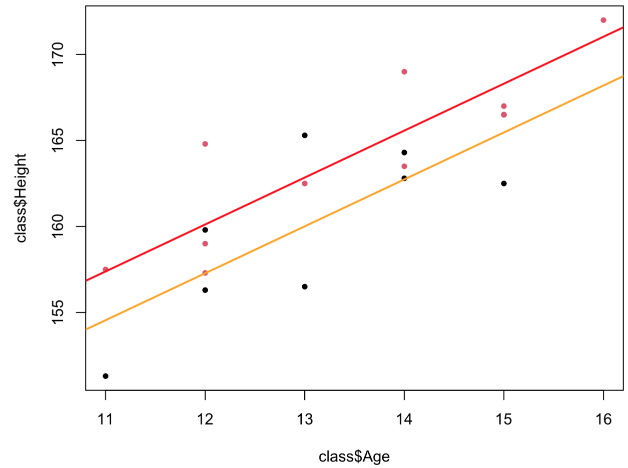
If this assumption is incorrect, it means that there is an *interaction* between the factors « age » and « gender », that is, the effect of age is different depending on the gender.

Interactions are modeled in R in the following way:

lm(formula = Height ~ Age + Gender + Age:Gender)

which is equivalent to

lm(formula = Height ~ Age * Gender)



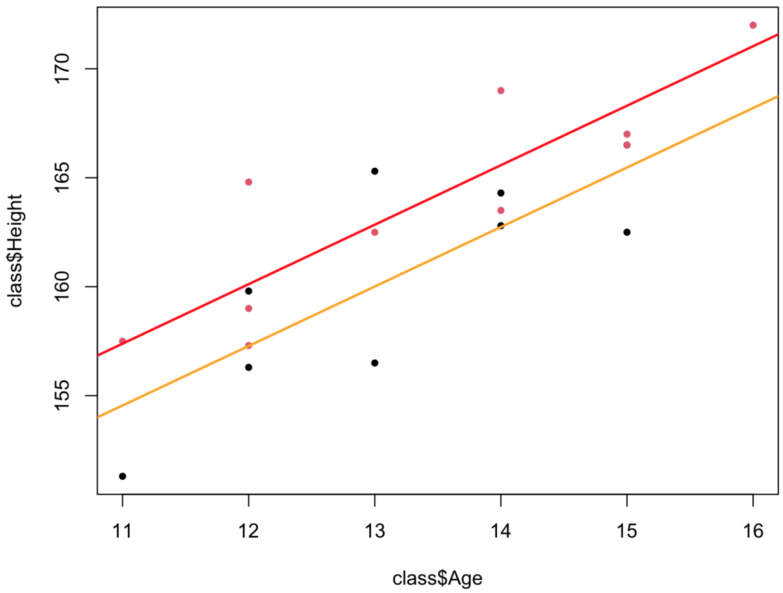
Coefficients with an interaction

```
Call:
lm(formula = Height ~ Age * Gender, data = class)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4429 -1.7844 -0.3648  1.3730  5.3571 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 122.1500     9.6409  12.670 2.05e-09 ***
Age           2.9071     0.7256   4.007  0.00114 ** 
GenderM       6.7443    12.4109   0.543  0.59483    
Age:GenderM  -0.2940     0.9285  -0.317  0.75585    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.862 on 15 degrees of freedom
Multiple R-squared:  0.7404, Adjusted R-squared:  0.6885 
F-statistic: 14.26 on 3 and 15 DF,  p-value: 0.0001152
```

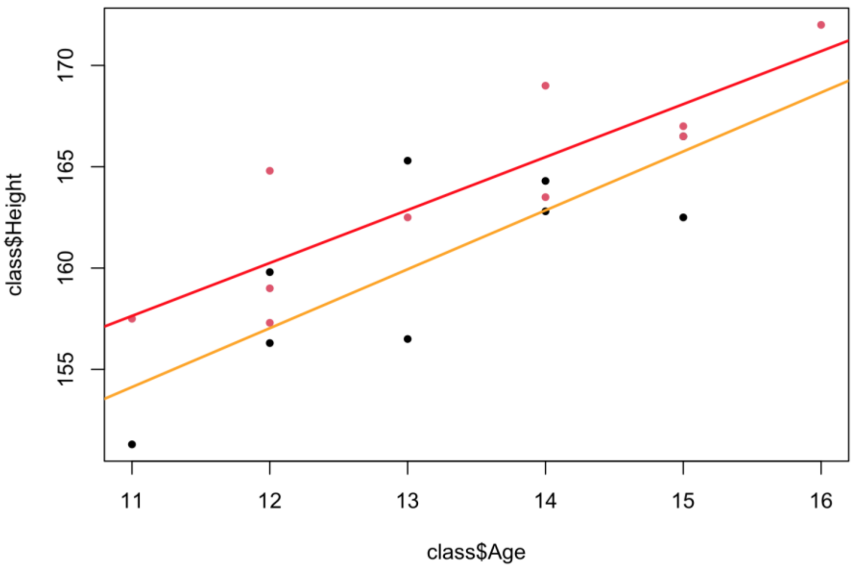
The coefficients can be interpreted as follows:

According to the model, the *height* is equal to

122.15 (the intercept)
plus 2.9071 times the person’s age
plus 6.7443, but only for males
-0.2940 times the person’s age, but only for males.



No interaction



With interaction

```
> model <- lm( Height ~ Age+Gender1, data=class)
> summary(model)

Call:
lm(formula = Height ~ Age + Gender1, data = class)

Residuals:
   Min     1Q Median     3Q    Max 
-3.483 -1.910 -0.319  1.326  5.317 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 127.3603     5.9587  21.374 3.43e-13 ***
Age           2.7276     0.4398   6.202 1.27e-05 ***
Gender1F     -2.8362     1.2797  -2.216   0.0415 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.78 on 16 degrees of freedom
Multiple R-squared:  0.7387,	Adjusted R-squared:  0.706 
F-statistic: 22.61 on 2 and 16 DF,  p-value: 2.176e-05
```

What if Males were the baseline ?

The two models are exactly the same; only the way we look at the coefficient changes.

```
> model <- lm( Height ~ Age+Gender, data=class)
> summary(model)

Call:
lm(formula = Height ~ Age + Gender, data = class)

Residuals:
   Min     1Q Median     3Q    Max 
-3.483 -1.910 -0.319  1.326  5.317 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 124.5241     5.8886  21.147 4.04e-13 ***
Age           2.7276     0.4398   6.202 1.27e-05 ***
GenderM       2.8362     1.2797   2.216   0.0415 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.78 on 16 degrees of freedom
Multiple R-squared:  0.7387,	Adjusted R-squared:  0.706 
F-statistic: 22.61 on 2 and 16 DF,  p-value: 2.176e-05
```

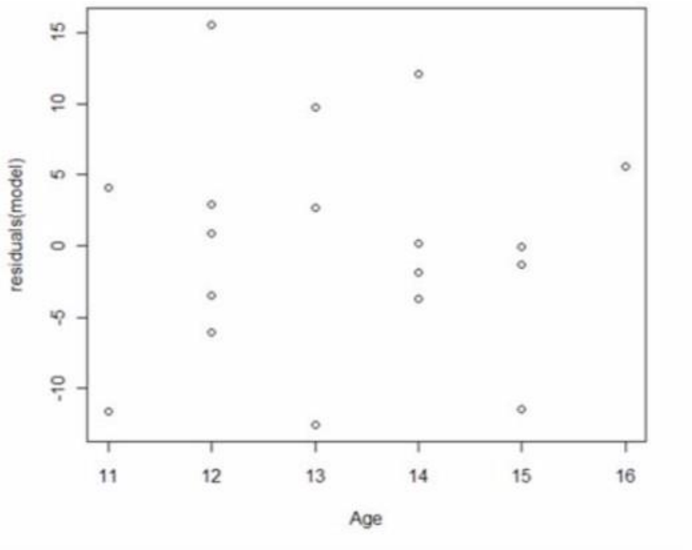
```
Gender1 <- relevel(Gender, ref="M")
```

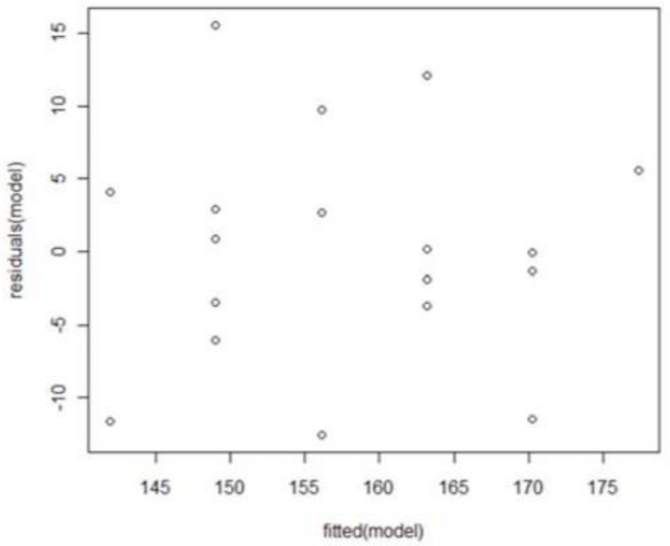
Diagnostic tools

It is always possible to fit a linear model and find a slope and intercept ... but it does not mean that the model is meaningful !

Examination of *residuals*: (which should show no obvious trend, since any systematic effect in the residuals should ideally be captured by the model):
– Normality
– Time effects
– Nonconstant variance – Curvature

Examination of *residuals*






```
plot( Age, residuals(model) )
```

```
plot( fitted(model), residuals(model) )
```

Works only for simple regression (only one variable on x axis)

Works also for multiple regression

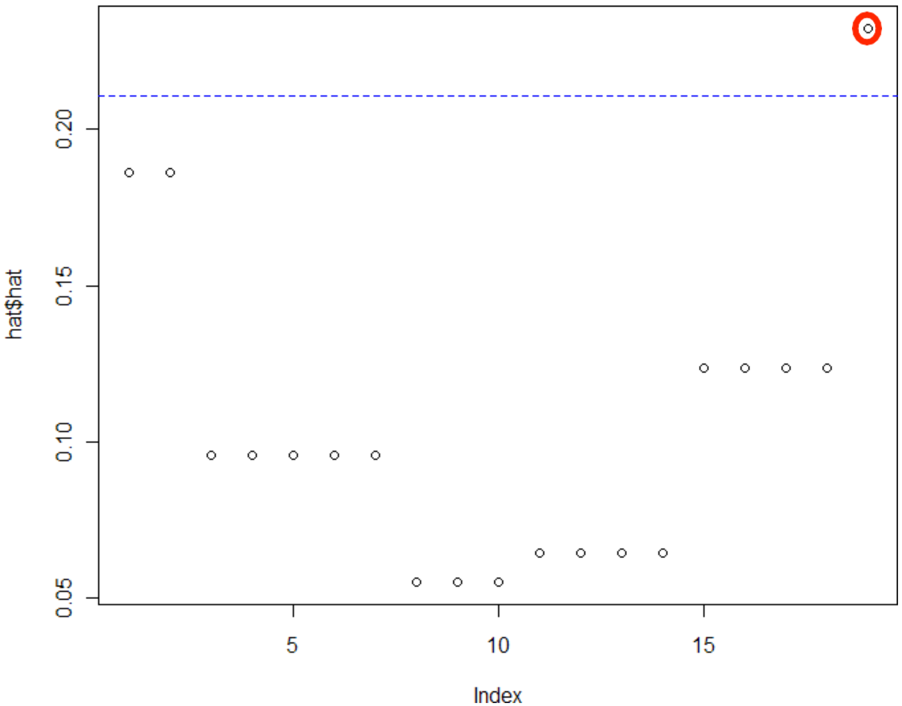
High leverage ('influential') points are far from the center, and have potentially greater influence

One way to assess points is through the *hat values* (obtained from the *hat matrix H*):

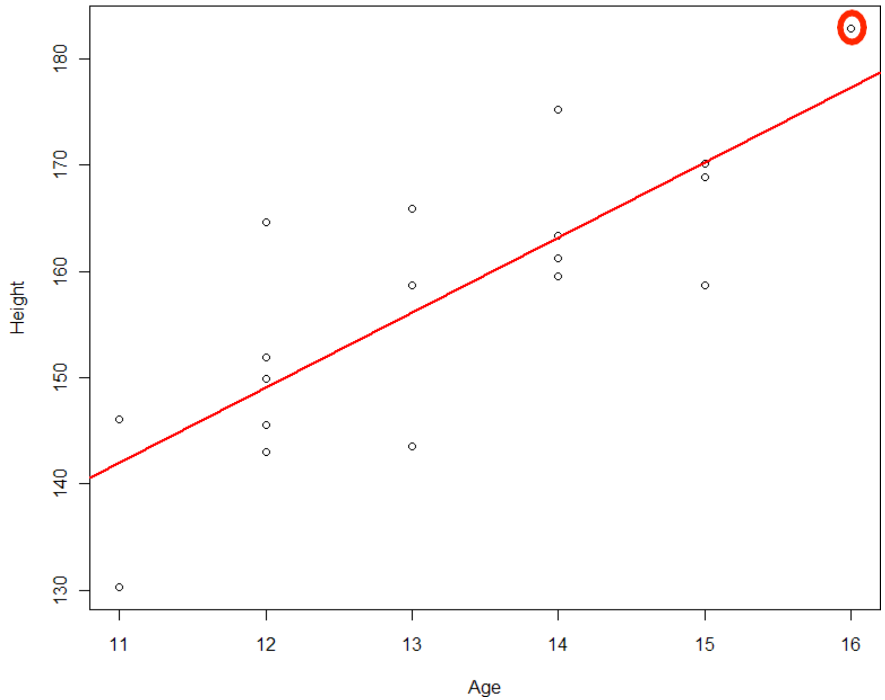
$$\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$$
$$h_i = \Sigma_j h_{ij_2}$$

Average value of h = number of coefficients/n (including the intercept) = p/n

Cutoff typically 2p/n or 3p/n

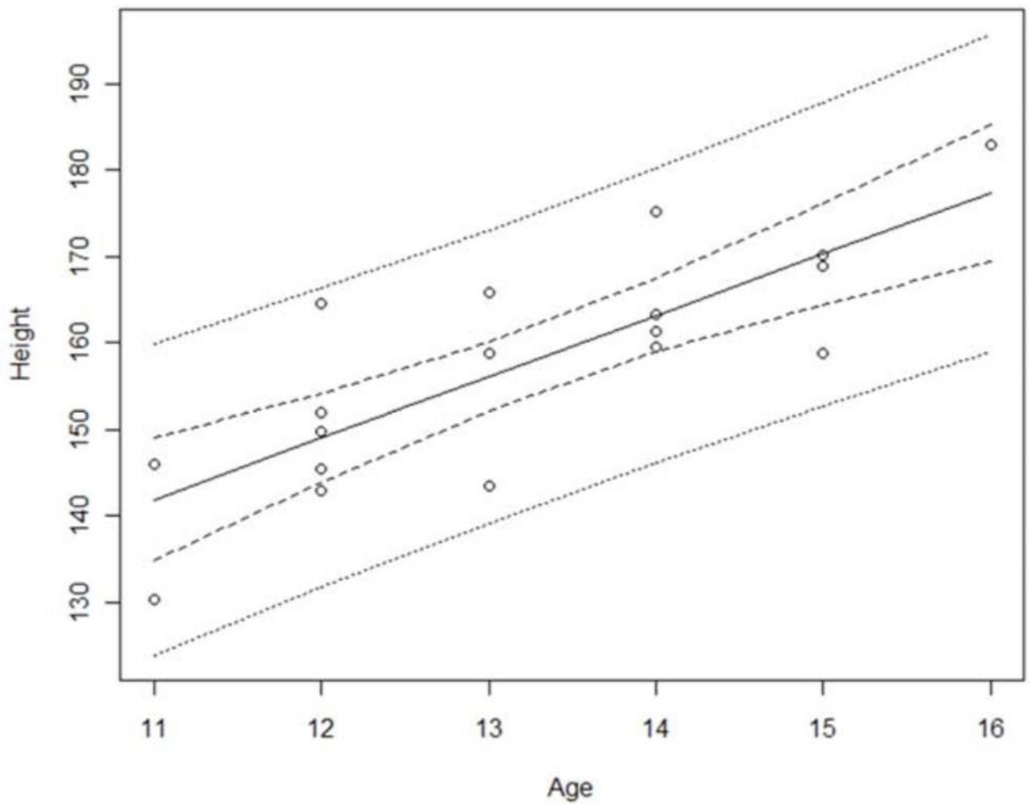


Hat values



Actual fit

```
>hat <- lm.influence( model )
>plot( hat$hat )
>abline(h=c(c(2,3)*2/19),lty=c(2,3),col=c("blue","red") )
```



Narrow bands: describe the uncertainly about the regression line

Wide bands: describe where most (95% by default) predictions would fall, assuming normality and constant variance.

In R: `?predict.lm`