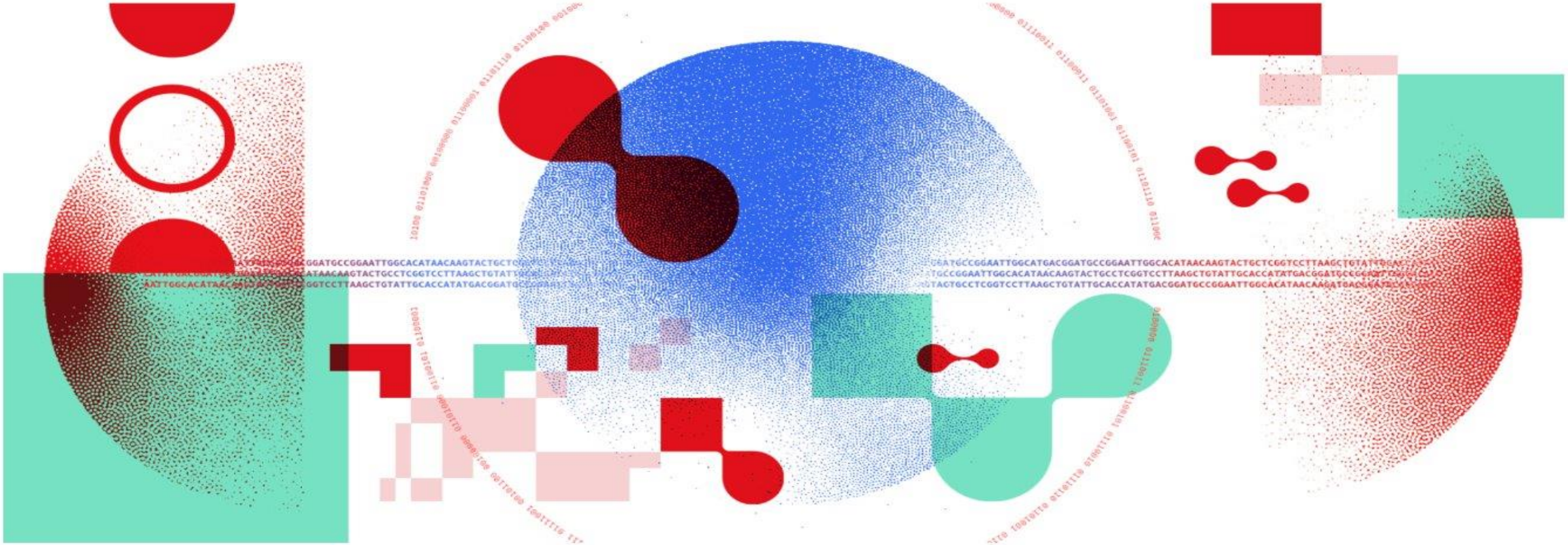


Advanced Statistical Modelling

Lausanne, September 2026

Paolo Angelino and Rachel Marcone

Simple and Multiple Linear Models

Plan

- Modelling- General rules
- Simple Linear Regression
- Multiple Linear Regression
- Assumptions
- Diagnostics

Modelling-General rules

- The idea of modelling is being able to find a good **representation** of the data in order maybe to understand what is **influencing** it or be able to **predict** future data.
- What is the best model ? Mostly it's a **trial and error**, and generally the patterns involved in the data are known (for example, linear, exponential, made of different steps).
- Start simple!
- Statistics help to understand if a model is better than another one but does not prove to be the only way to look at your data.
- Modelling is an approximation of the data and will never be perfect.

Statistical Modelling

- Statistical modelling is a set of **equations** that are solved, along with a set of random variables that follow certain **assumptions.**

- **Which equations ? Which assumptions ?**

Variables and questions

- **Dependent variables (or responses):** variables we want to describe, understand, explain, model, predict

- **Explanatory variables**
- **(or independent variables or predictors or covariates):** variables we use to explain, to describe or to predict the dependent variable(s). These are usually linked to the biological question you ask (for example : does gender have an effect on a drug response)

Modelling overview

- Many models will be written in the following formula
- $Y = f(X) + error$

- Sometimes a transformation of the Y variable is needed for a relationship so models can be also of the form
- $g(Y) = f(X) + error$

- Y is the dependent variable
- And here X is the only explanatory variables but there could be many
- $Y = f(X_1) + f(X_2) + f(X_3)+\ldots+ error$
- The error term is what is of interest. We will try to **minimize** the error term and for that we will have to make **assumptions** on the error term.

Modelling overview in R

- The generic form of a model in R is given by
  - Response ~ predictors
- + to **add** more variables
- - to **leave out** variables
- : to indicate **interaction** between two terms
- * to include both interaction and the terms
- ^n adds all terms including interaction up to degree n
- I() treats what is in the parenthesis as mathematical expression

Modelling overview in R

- The generic form of a model in R is given by
  - Response ~ predictors
- + to **add** more variables
- - to **leave out** variables
- : to indicate **interaction** between two terms
- * to include both interaction and the terms
- ^n adds all terms including interaction up to degree n
- I() treats what is in the parenthesis as mathematical expression
  - **Income~ age^2 + I(weight/height^2) +gender**

Model formulas in R

- A simple formula in R can be for example
  - **`yvar ~ xvar1 + xvar2 + xvar3`**
- Can read ~ as “described (or modeled) by”.
- We could write a model (algebraically) as
  - Y = b0 + b1 x1 + b2 x2 + b3 x3 + error

Model formulas in R

- By default, an intercept is included in the model – you don't have to include a term in the model formula

- If you want to leave the intercept out:
    - `yvar ~ -1 + xvar1 + xvar2 + xvar3`

- Can read ~ as "described (or modeled) by".

- We could write a model (algebraically) as
    - Y = b1 x1 + b2 x2 + b3 x3 + error

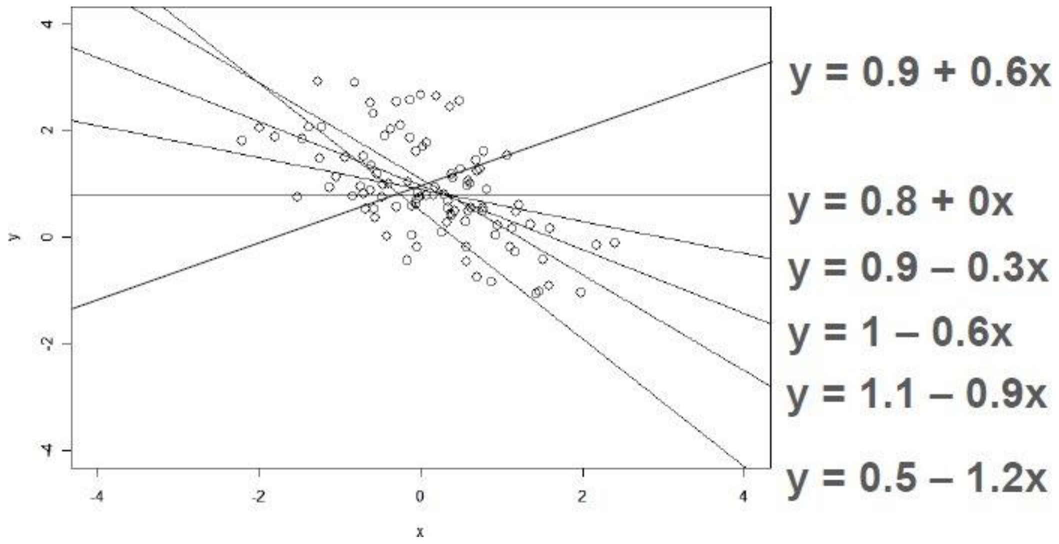
Model formulas in R

- If you only want the intercept, this is called the null model, or the estimation of the overall mean, or the grand mean
    - `yvar ~ 1`
- We could write a model (algebraically) as
    - Y = b0+ error

Linear regression

Linear Regression

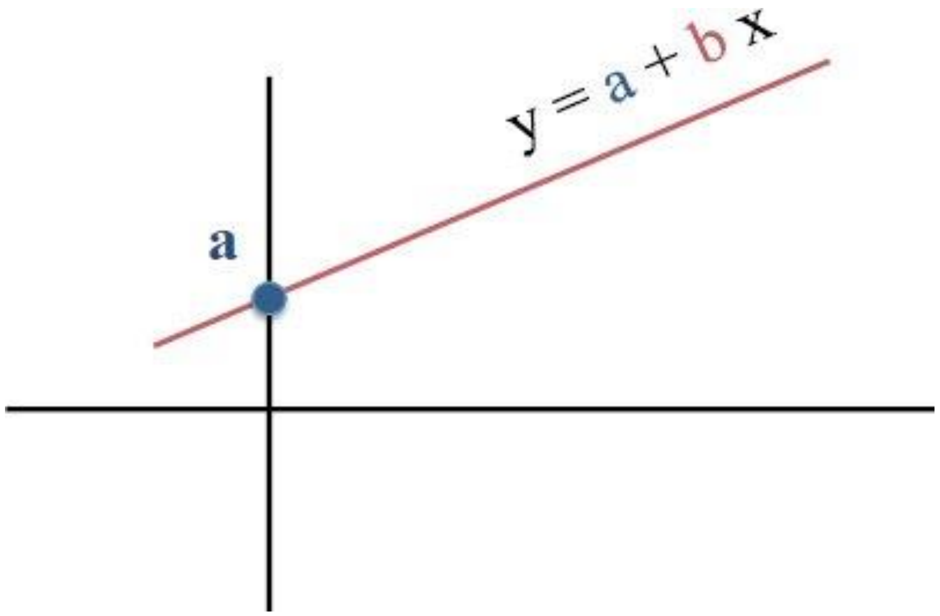
- Linear Regression = fit a line that summarises the relationship between x and y. Which is the best line ? What criteria ?



Linear Regression

- The equation for a line to predict y knowing x is given by

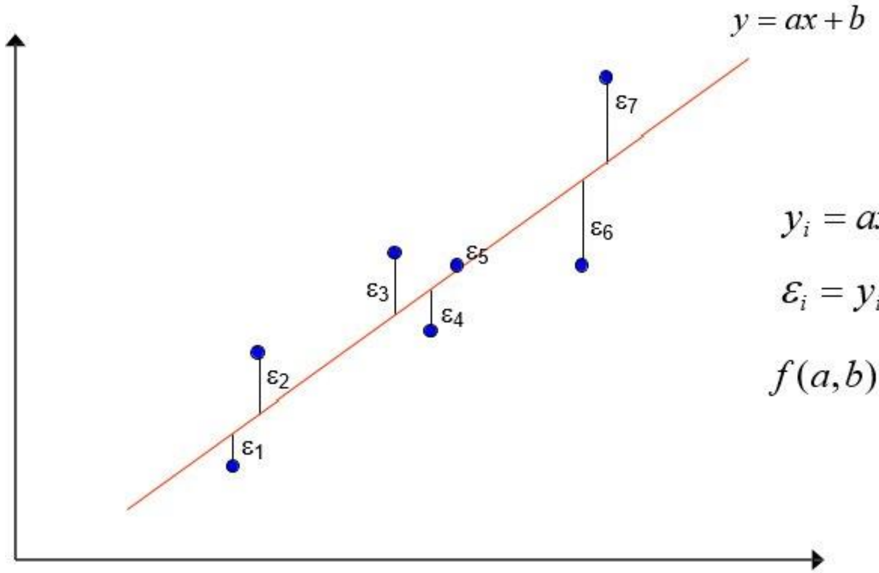
- Y = a +bx where a is called the intercept and b the slope



Least squares fitting (LS)

- One possibility-not the only one!
- The least square fitting finds the straight line with the smallest sum of squares of vertical errors

Least-square fitting



Regression line such that:

$$\sum_i \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + ..$$

minimum

$$y_i = ax_i + b + \varepsilon_i$$

$$\varepsilon_i = y_i - (ax_i + b)$$

$$f(a,b) = \sum_i \varepsilon_i^2 = \sum_i \left[y_i - (ax_i + b)\right]^2$$

$$\boxed{\begin{array}{l} \partial f(a,b)/\partial a = 0 \\ \partial f(a,b)/\partial b = 0 \end{array}}$$

Mathematically

Formalization and extension of linear regression

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

$i = 1, \cdots, n$

Y represents **one** data point

Y_i : response (known)

β_0, β_1 : model parameters (estimated)

X_i : predictor (known)

ε_i : error term $\sim N(0, \sigma^2)$ (estimated)

Minimizing $\sum_i \varepsilon_i^2$ yields b_0 and b_1 estimators of β_0 and β_1

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Error terms or Model residuals

- Are the distances between data points and the expected values based on the model (equation with fitted parameters)
- Model residuals represents the part of **variability** in the data the model was **unable to capture**
- Errors in least square fits are normally distributed with mean 0 and **constant variance**.

Warning

- 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 the residuals is important
  - No trend in the residuals – any systemic effect should be captured by the model
  - Normality
  - Constant variance

Residuals

- The residuals are the error made when predicting the data using the regression line, therefore they are given by
  - Error = Ymeasured-Ypredicted = Ymeasured – (a+bX)
- The regression equation has the property that :
- The sum of the residuals is 0 which is the same as saying the mean of the residuals is 0

Multiple Linear regression

Linear models (general case)

p parameter linear model

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i} \qquad i = 1, \cdots, n$$

or $$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i \qquad \text{with} \qquad X_{i0} \equiv 1$$

Y_i response (e.g. expression of a gene)

X_{ik} predictor variables (e.g. dose of drug [continuous], or KO vs wt)

β_k model parameter (measurement of magnitude of effect associated to predictor variable)

ε_i error term (measurement of departure from ideal case)

Linear models: matrix form

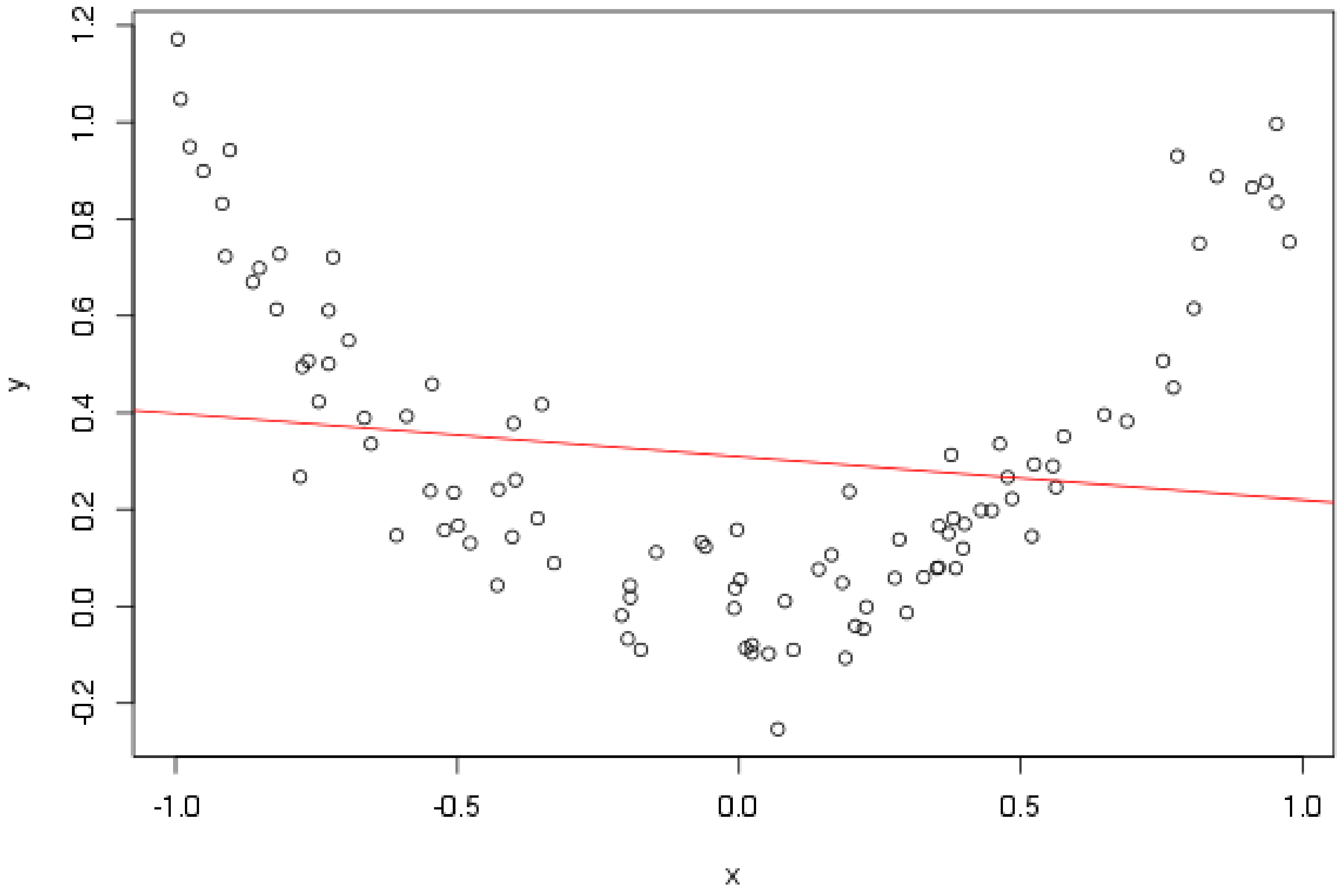
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$$

is equivalent to

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p-1} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

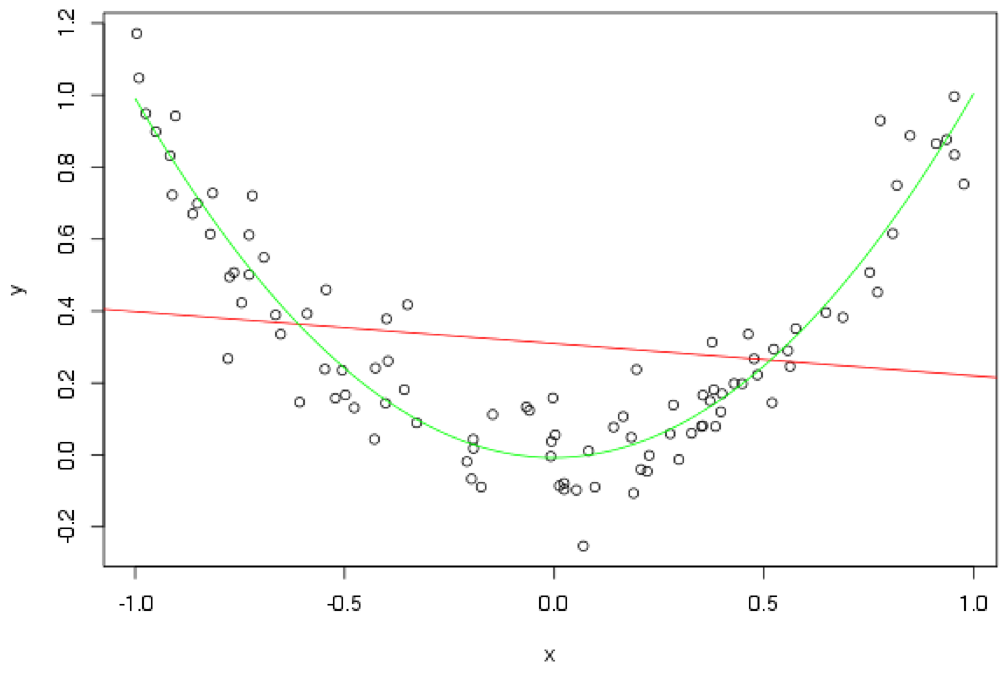
What if the association is not linear ?



What if the association is not linear ?

Use a polynomial regression

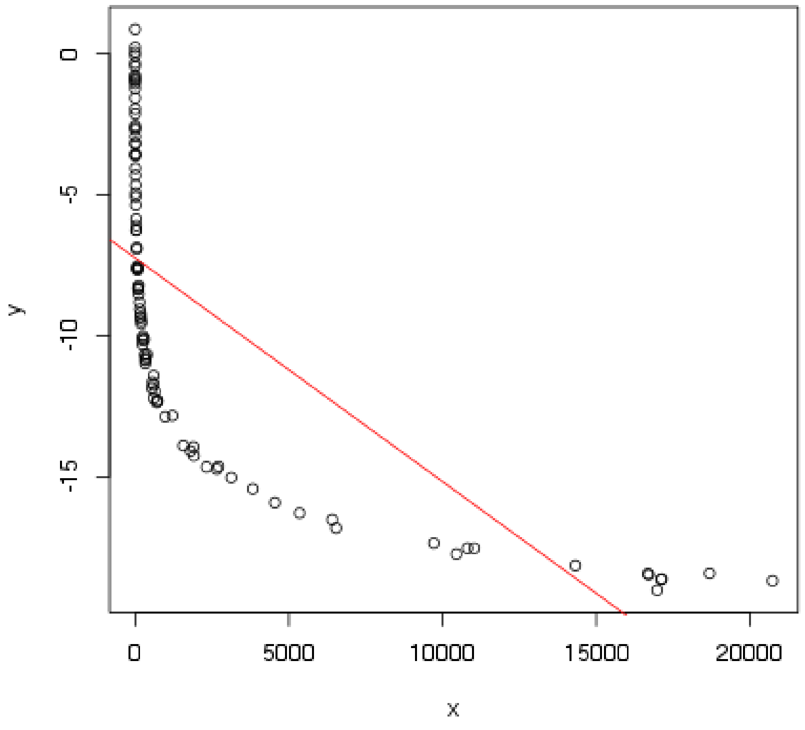
$$y = b_0 + b_1 x + b_2 x^2$$

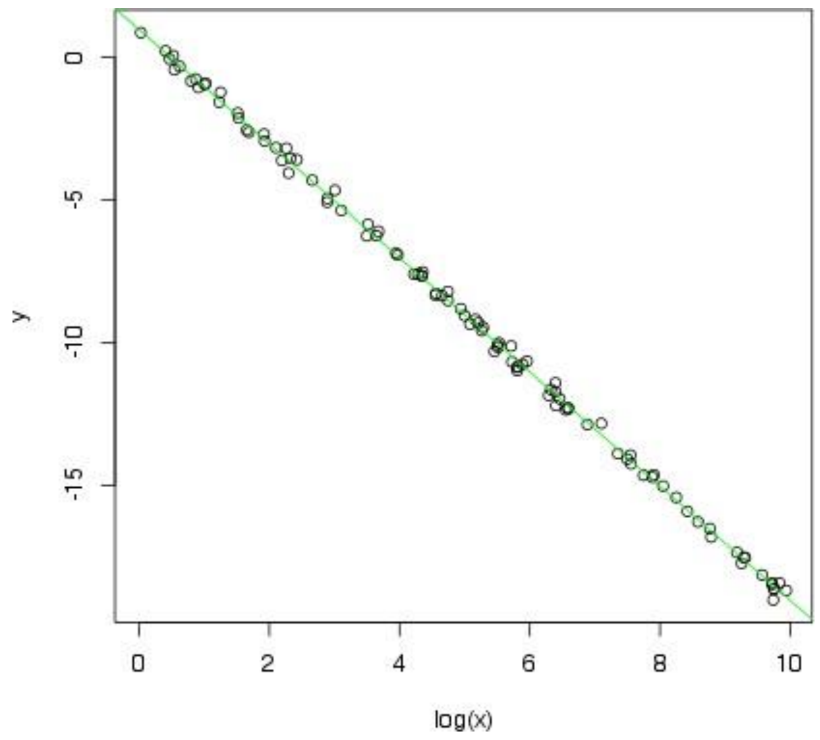


What if the association is not linear ?

Consider transforming the data (log)

log(y) = a + b x





Linear or not linear ?

Linearity is about the model parameters

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{ip-1} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

$$Y_i = \beta \sin X_i + \varepsilon_i$$

Linear in βs

$$Y_i = \beta_0 + \log(\beta_1 X_{i1} + \beta_2 X_{i2}) + \beta_3 X_{i3} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 \exp(\beta_2 X_i + \beta_3) + \varepsilon_i$$

Not linear in βs

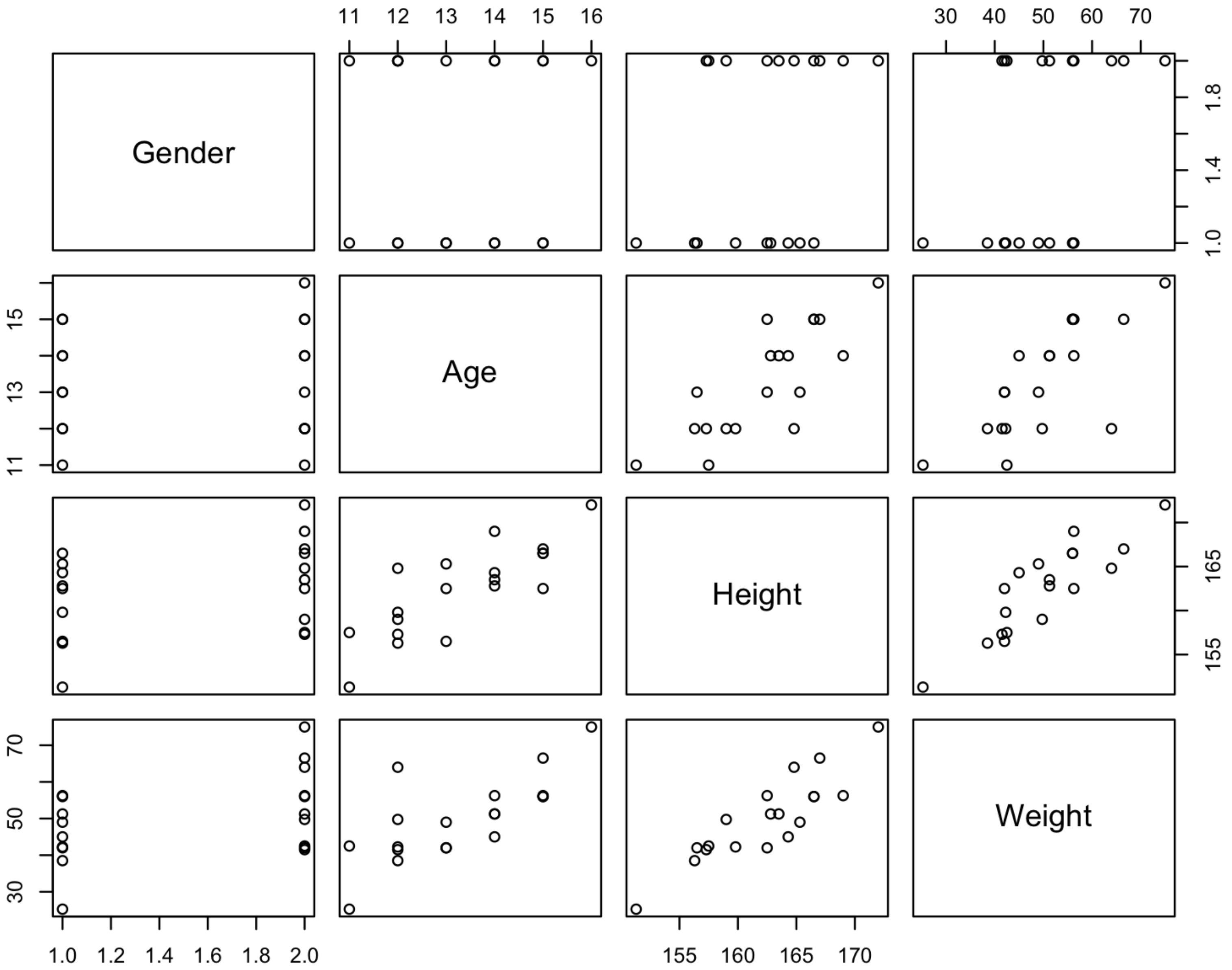
Concrete example

The CLASS dataset

```
> class
      Name Gender Age Height Weight
1    JOYCE      F  11  151.3  25.25
2   THOMAS      M  11  157.5  42.50
3    JAMES      M  12  157.3  41.50
4     JANE      F  12  159.8  42.25
5     JOHN      M  12  159.0  49.75
6   LOUISE      F  12  156.3  38.50
7   ROBERT      M  12  164.8  64.00
8    ALICE      F  13  156.5  42.00
9  BARBARA      F  13  165.3  49.00
10 JEFFREY      M  13  162.5  42.00
11   CAROL      F  14  162.8  51.25
12   HENRY      M  14  163.5  51.25
13  ALFRED      M  14  169.0  56.25
14    JUDY      F  14  164.3  45.00
15   JANET      F  15  162.5  56.25
16    MARY      F  15  166.5  56.00
17  RONALD      M  15  167.0  66.50
18 WILLIAM      M  15  166.5  56.00
19  PHILIP      M  16  172.0  75.00
```

The CLASS dataset

```
> summary(class)
     Name              Gender               Age            Height     
 Length:19          Length:19          Min.   :11.00   Min.   :151.3  
 Class :character   Class :character   1st Qu.:12.00   1st Qu.:158.2  
 Mode  :character   Mode  :character   Median :13.00   Median :162.8  
                                       Mean   :13.32   Mean   :162.3  
                                       3rd Qu.:14.50   3rd Qu.:165.9  
                                       Max.   :16.00   Max.   :172.0  
     Weight     
 Min.   :25.25  
 1st Qu.:42.12  
 Median :49.75  
 Mean   :50.01  
 3rd Qu.:56.12  
 Max.   :75.00  
> pairs(class[,-1])
```



Gender
Age
Height
Weight
11 12 13 14 15 16
30 40 50 60 70
1.0 1.4 1.8
11 13 15
155 165
30 50 70
1.0 1.2 1.4 1.6 1.8 2.0
155 160 165 170

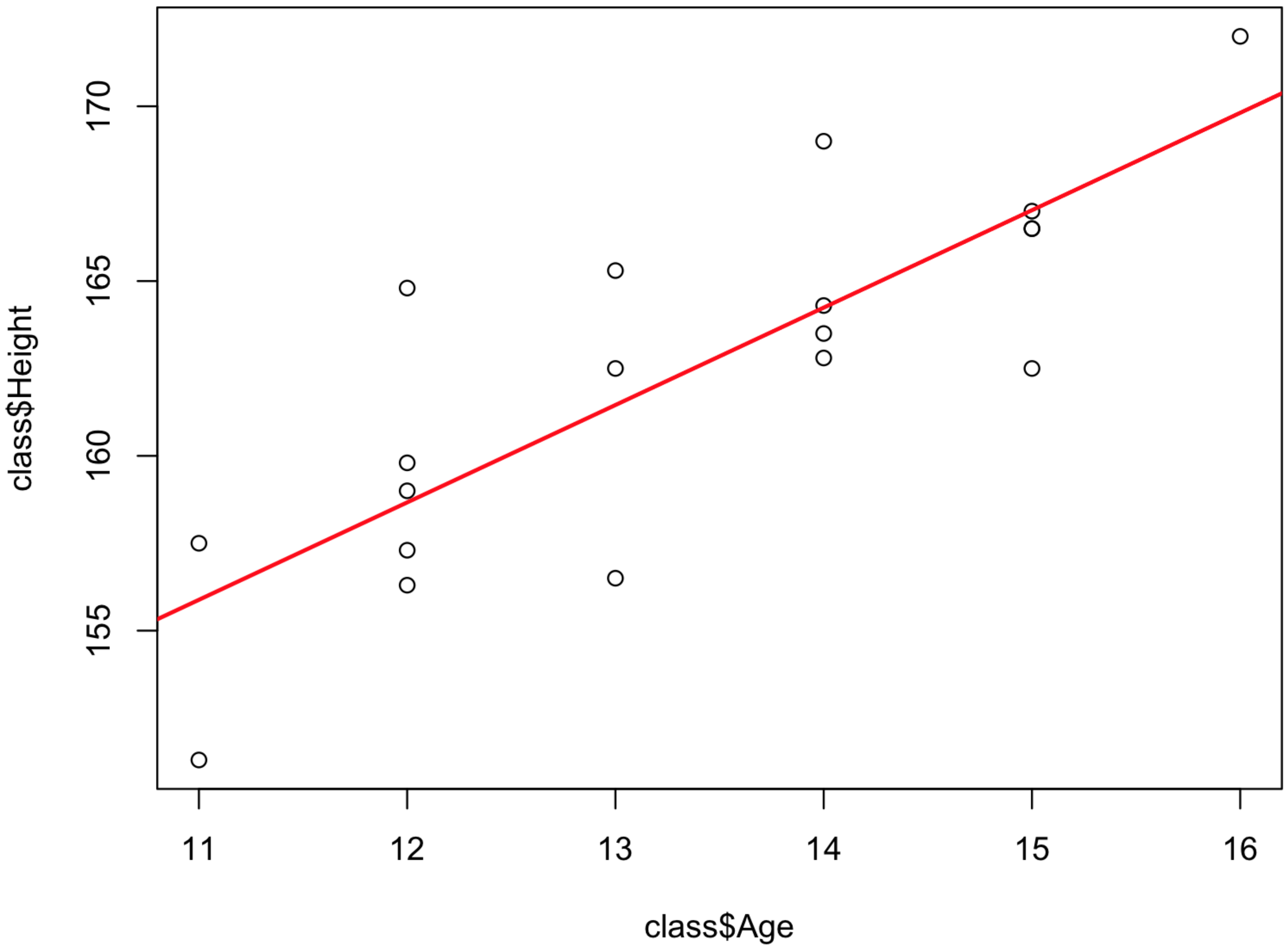
Fitting a linear model

```
> lm( Height ~ Age, data=class)

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

Coefficients:
(Intercept)          Age
    125.224        2.787

> model <- lm( Height ~ Age, data=class)
> model

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

Coefficients:
(Intercept)          Age
    125.224        2.787
```

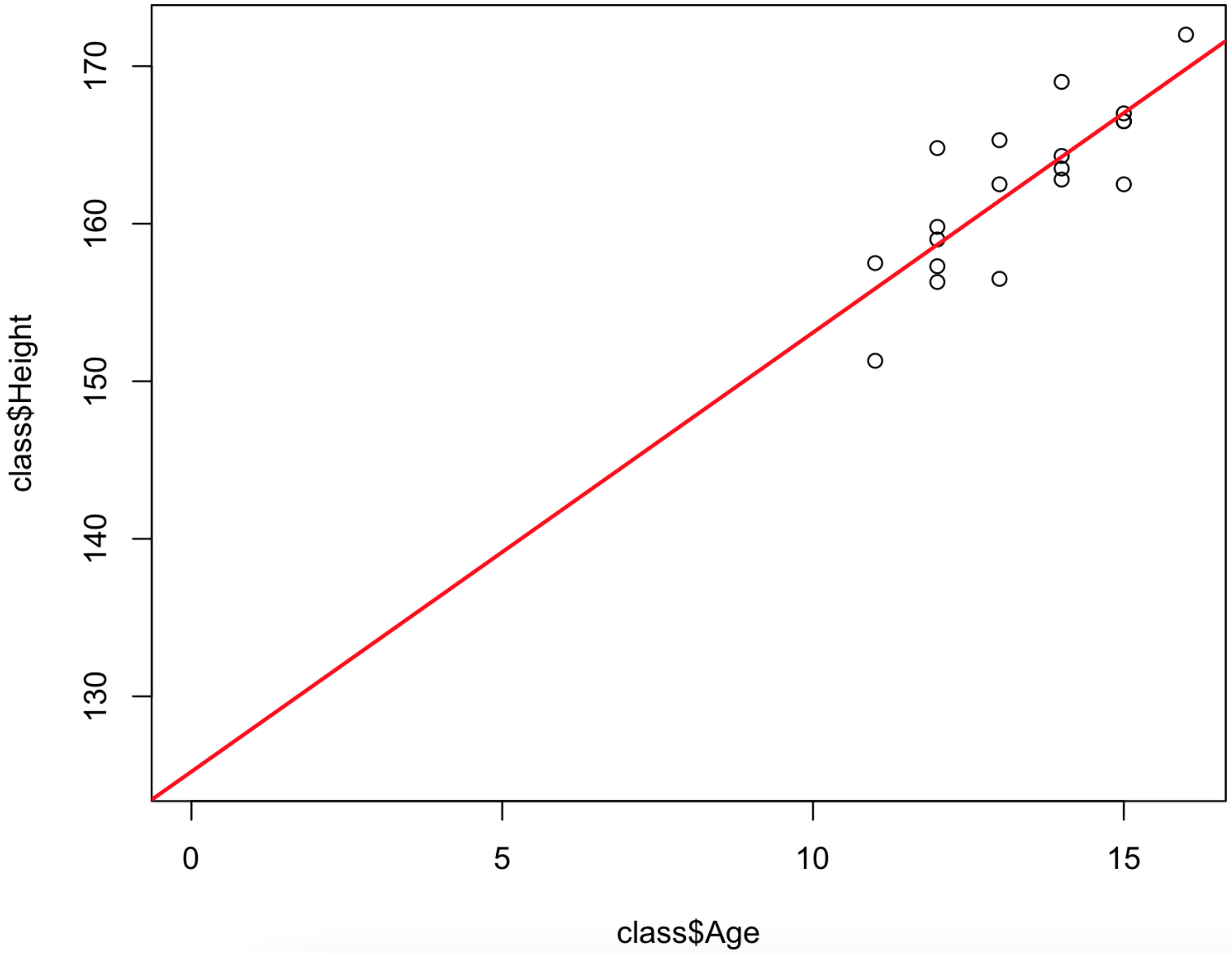
Model:

Height = 125.224 + 2.787x Age

```
> plot( class$Age, class$Height)
> abline(model, col="red", lwd=2)
```



```
> plot(class$Age, class$Height,
  xlim=range(0,Age),
  ylim=range(coef(model)[1], Height))
> abline(model, col="red", lwd=2)
```



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

```
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max 
-4.957 -1.407 -0.031  1.374  6.130 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 17 degrees of freedom
Multiple R-squared:  0.6584, Adjusted R-squared:  0.6383 
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

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

```
> summary(model)

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

Residuals:
   Min      1Q  Median      3Q     Max 
-4.957 -1.407 -0.031  1.374  6.130 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 17 degrees of freedom
Multiple R-squared:  0.6584, Adjusted R-squared:  0.6383 
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

Function call

Example of a summary of lm command

```
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max 
-4.957 -1.407 -0.031  1.374  6.130 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 17 degrees of freedom
Multiple R-squared:  0.6584, Adjusted R-squared:  0.6383 
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

Residuals

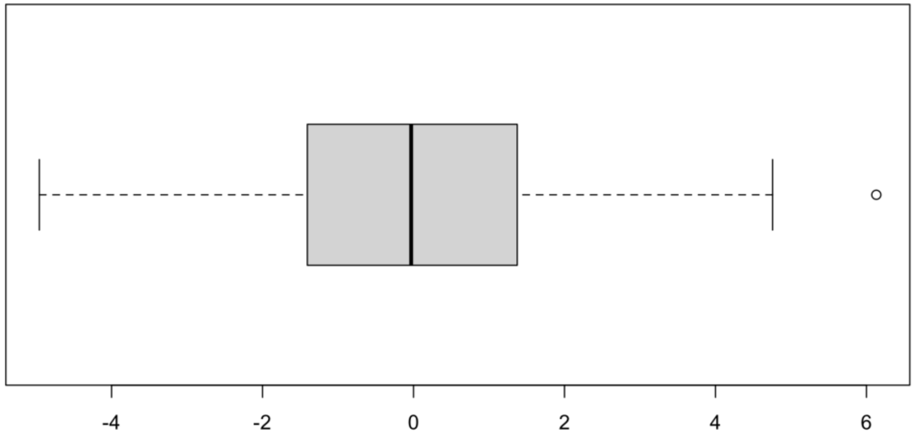
Five-number summary of the residuals equivalent to

```
> fivenum( residuals( model ) )
```

```
          8          11          17           4           7
-4.95669291 -1.40669291 -0.03097113  1.37401575  6.13044619
```

or, graphically, using a boxplot:

```
>boxplot( residuals ( model), horizontal=T)
```



Coefficients and p-values

```
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max 
-4.957 -1.407 -0.031  1.374  6.130 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 17 degrees of freedom
Multiple R-squared:  0.6584, Adjusted R-squared:  0.6383 
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

Coefficients

These statistical tests tell us if the parameters are significantly different from 0.

****It is not interesting for the intercept, but usually interesting for the slope.**

Estimate and Std. Error are used for hypothesis testing

T-value = Estimate / Std. Error

This assumes that the residuals follow a normal distribution!

RSE

```
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max 
-4.957 -1.407 -0.031  1.374  6.130 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 17 degrees of freedom
Multiple R-squared:  0.6584, Adjusted R-squared:  0.6383 
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

RSE-formula

$$\text{RSE} = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- y_i = actual observed value
- $\hat{y}_i$ = predicted value from the regression
- n = number of observations
- p = number of parameters in the model (including the intercept)
- $(n - p)$ = degrees of freedom

RSE (Residual Standard Error) and degrees of freedom

The number of *degrees of freedom* indicates the number of independant pieces of data that are available to estimate the error

While we have 19 residuals here, they are not all independent: for example, the last one is constrained because the sum of all residuals must be 0.

The number of DF

total observations – number of parameters estimated

Two parameters are estimated (intercept + coefficient), so 19-2 = 17

R-squared

```
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max 
-4.957 -1.407 -0.031  1.374  6.130 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 17 degrees of freedom
Multiple R-squared:  0.6584, Adjusted R-squared:  0.6383 
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

Multiple and adjusted R-squared

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

if $R^2=1$, the data fits perfectly on a straight line, and the model explains all the variance

Multiple and adjusted R-squared

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

if R^2=1, the data fits perfectly on a straight line, and the model explains all the variance

In the case of simple regression, it is equal to the square of the correlation coefficient between the two variables:

```
> summary(model)$r.squared
[1] 0.6584257
> cor(class$Age,class$Height)^2
[1] 0.6584257
```

The Adjusted R-squared is similar to R-squared, but it takes into account the number of variables in the model (we will come back to this later).

F statistics

```
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max 
-4.957 -1.407 -0.031  1.374  6.130 

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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 17 degrees of freedom
Multiple R-squared:  0.6584,	Adjusted R-squared:  0.6383 
F-statistic: 32.77 on 1 and 17 DF,  p-value: 2.48e-05
```

F test

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

It calculates the F value which is given by the variation explained by our model divided by the variation that remains.

Mathematically : $\frac{SS(mean) - SS(fit)/(pfit - pmean)}{SS(fit)/(n - pfit)}$

Pfit= number of parameters in the fit (2 parameters)
Pmean = number of parameters in the mean line (1 parameter)
Note: With only one variable, it provides *exactly* the same result as the t-test for the significance of the coefficient of this variable.

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 onemultiple 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 and adjusted R-squared

```
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 R^2.

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

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
```

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 variable

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

Height = b0 + b1 Age + **b2 « Gender »** + 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.

$$\textbf{Gender} = \begin{cases} 1 \text{ if Female} \\ 2 \text{ if Male} \end{cases}$$

and the model can be interpreted as follows:

- b_0 is the baseline for height among women (at Age = 0)
- $\mathbf{b_2}$ represent the increase/decrease of this baseline for men.

- .

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

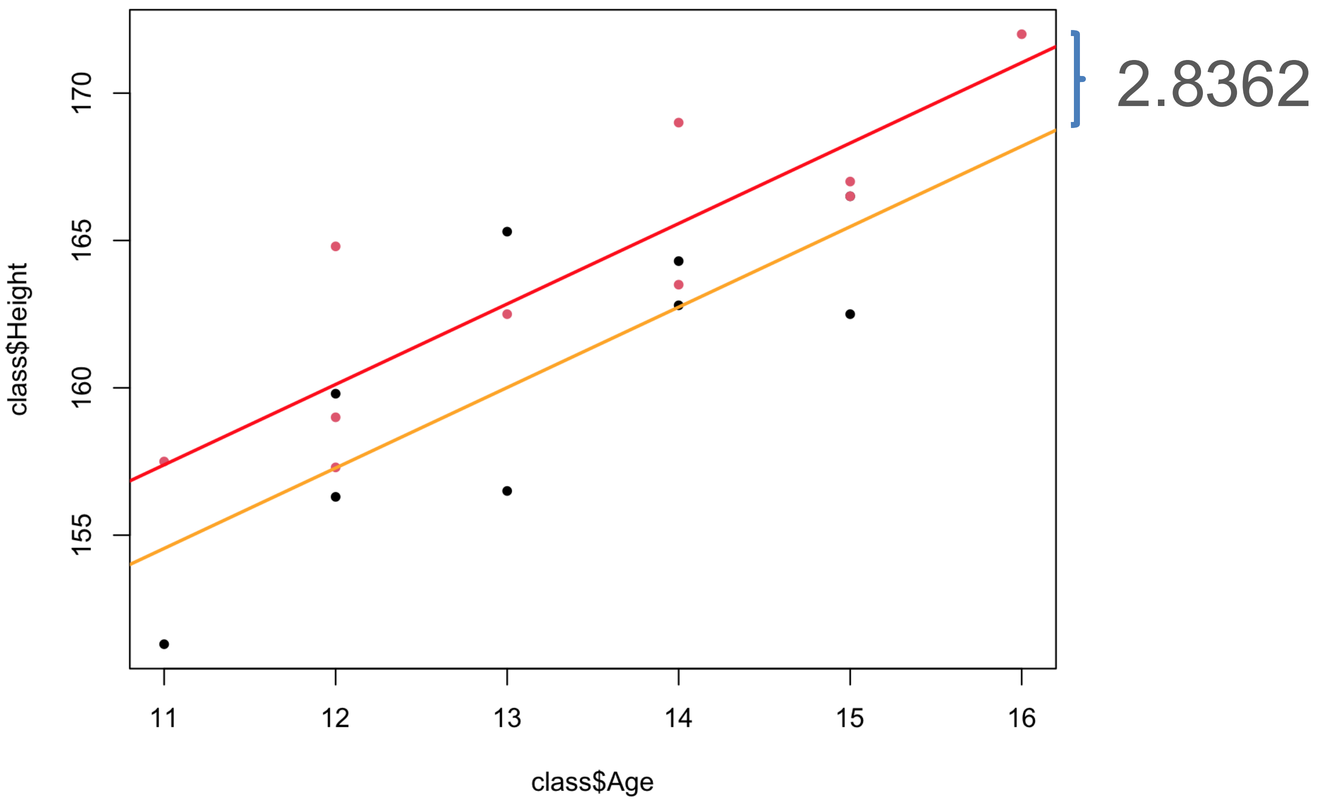
The factor GenderM corresponds to the difference in baseline for Males compared to females

Graphical interpretation

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 ?

Interactions

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.

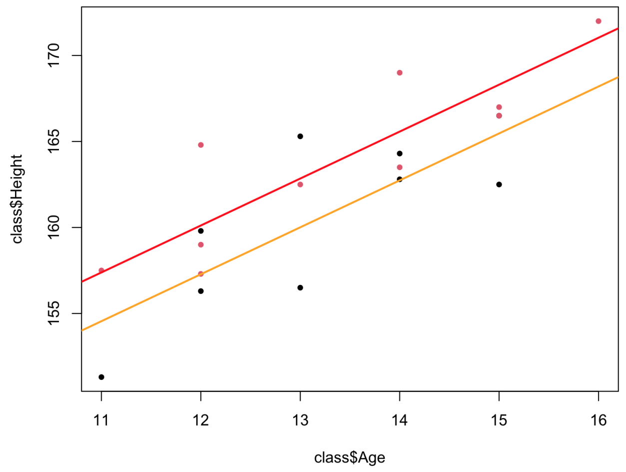
If this assumption is incorrect, it means that there is an *interaction* between the factors « age » and « gender », that is, the effect of age is different depending on the gender.

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

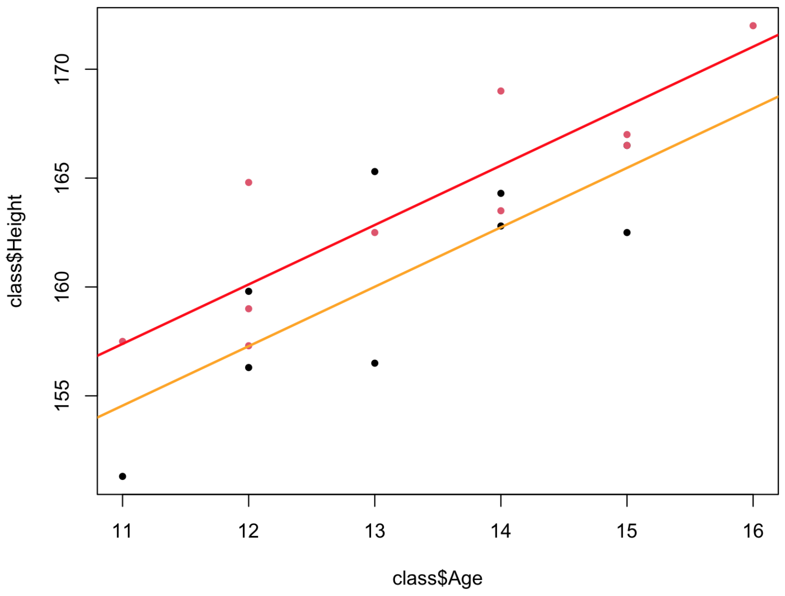
```
Call:
lm(formula = Height ~ Age * Gender, data = class)

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:

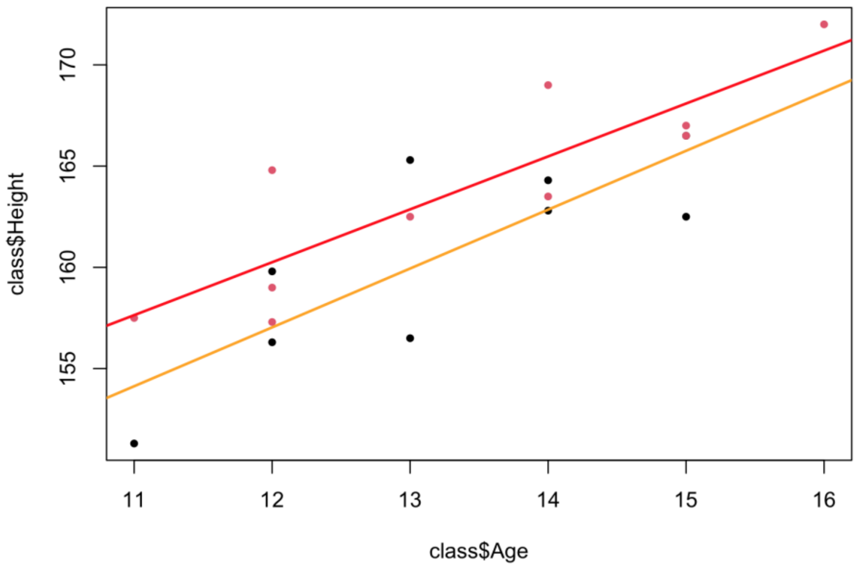
According to the model, the *height* is equal to

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.

Different slopes



No interaction



With interaction

What if Males were the baseline ?

```
> 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
```

```
> 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
```

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

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

Diagnostics

Basic model checking

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” or “optimal” !

Model checking **questions**:

- How good is the fitting , could it be improved ?
- Are there **outliers**, which might “disturb” during fitting ?
- Are there points that have a **high impact** and might decrease the model quality ?

- Does the the model fit look “perfect”: the residuals are “normal” (“Gaussian”) and have “constant variance” ? And are independent from each other ?

technically $E_i \sim$ i.i.d $N(0,$ sigma constant | i)

Note: the statistical tests (p-values) and confidence intervals are calculated using this assumptions, they are unreliable if this is not at least approximately satisfied.

Basic model checking

Examination of ***Residuals***:

- If they show a **pattern** => maybe can improve the model, there is still a systematic trend that could be captured by a better model

- If they have **variable variance**:

Is the model missing something compared to reality (is miss-specified): another explanatory variable ?, a data transformation?,
Or maybe there are some outliers that impact the parameter estimation?

- If they are **NOT normally** distributed: same questions

- If they are **NOT independent**: were the data collected in a “good way” ?

- If there are **Outliers**: which points (is OK? eliminate ? Can be verified ? Is it reproducible or one time only)

Residuals

Types of *Residuals*:

- **Raw residuals** Ri = Observed - Fitted (Expected) = Y - Y_i

- Rescaled (specifically to each data point) to have expected sd = 1 :
  Studentized Residuals
- Should follow about a t-distribution (resp. approx. N(0,1))

Hat Values and diagnostics

Ypredicted = X β
X' Y = X' X β => because X' (Y-X β) = 0 as these planes are orthogonal
$(X' X)^{-1}X'Y = \beta$
Ypredicted = $X (X' X)^{-1}X' Y = H Y$

This is the H matrix is the way to transform the measured values to the predicted values and is a very interesting matrix to look at.
The diagonal values are a good indication of influence of the points in the regression

Hat values diagnostics

Average value of h = number of coefficients (including the intercept)/number of observations= p/n
In a simple case it would be 2/n

A cutoff of 2*p/n or 3*p/n has been proposed in the litterature for influential points. *

*Belsley, Kuh, and Welsch (1980)

Cooks distance

For a point i, Cook's Distance D_i is calculated as:

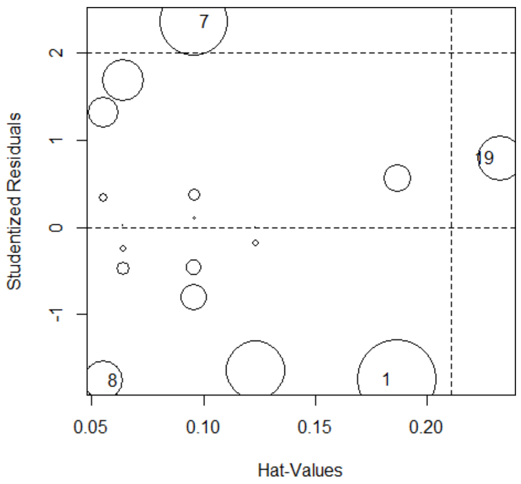
$$D_i = \frac{(r_i^2)}{p \cdot MSE} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}$$

Where:

- r_i: Studentized residual for observation i
- h_{ii}: Leverage of observation i
- p: Number of parameters in the model (including the intercept)
- MSE: Mean squared error of the model

influencePlot

```
library(car)

influencePlot(model,
xlab="Hat-Values",
ylab="Studentized
Residuals")
```

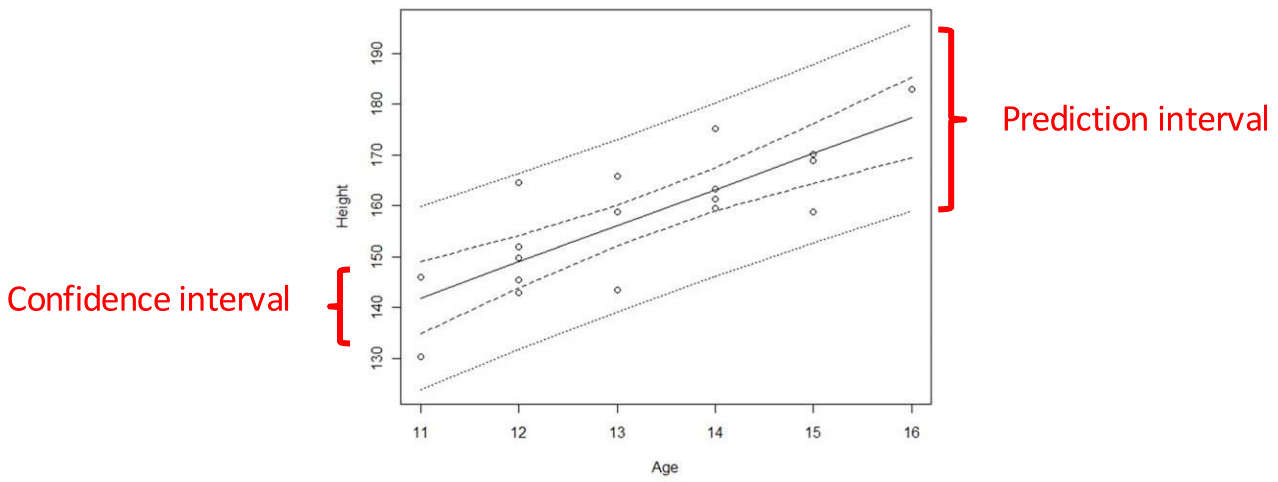


Description
This function creates a “bubble” plot of Studentized residuals versus hat values, areas of the circles proportional to the value “Cook's distance”.

Vertical reference lines are drawn at twice and three times the average hat value, horizontal reference lines at -2, 0, and 2 on the Studentized-residual scale.

Value
If points are identified, returns a data frame with the hat values, Studentized residuals and Cook's distance of the identified points. If no points are identified, nothing is returned. This function is primarily used for its side-effect of drawing a plot.

Confidence bands



Narrow bands: describe the uncertainty about the regression line

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

```
predict.lm(model, newdata=data.frame(Age=new_age), interval="confidence")
predict.lm(model, newdata=data.frame(Age=new_age), interval="prediction")
```