The Linear Model

Livio Finos

Contents

1 Outline

1.1 Outline

- Covariance and Correlation
- Simple Linear Model
- Analysis of the residuals
- 2 sample
- Multiple Linear Model
- Anova
- Interaction terms

1.2 The Age vs Reaction Time Dataset

The reaction time of these subjects was tested by having them grab a meter stick after it was released by the tester. The number of centimeters that the meter stick dropped before being caught is a direct measure of the person's response time.

The values of `Age` are in years. The `Gender` is coded as `F` for female and `M` for male. The values of `Reaction.Time` are in centimeters.

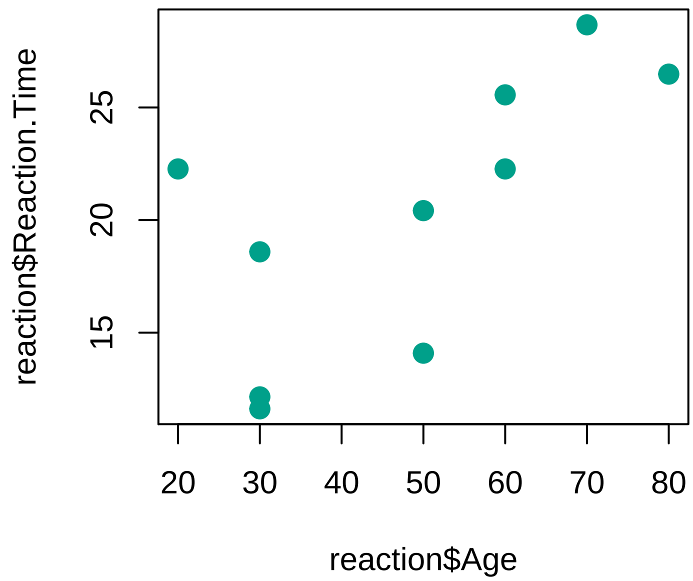
(data are fictitious)

To read the data

```
data(reaction,package = "flip")
# or download it from: https://github.com/livioivil/flip/tree/master/data
# str (reaction)
```

We plot the data

```
plot(x=reaction$Age,y=reaction$Reaction.Time,pch=20,col=2,cex=2)
```



2 Measures of Dependence and the Simple linear model

2.1 Measuring the dependence

we define:

- $X = Age$
- $Y = Reaction.Time$

We review some famous index to measure the (linear) dependence among two variables

2.2 Covariance and Variance

Covariance between X and Y:
$\sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$

- values between $-\infty$ and ∞
- $\sigma_{xy} \approx 0$: there is no dependency between X and Y
- $\sigma_{xy} >> (<<) 0$: there is a strong positive (negative) dependency between X and Y

Variance of X (= covariance between X and X):

$\sigma_{xx} = \sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$

Standard Deviation of X:

$\sigma_{xx} = \sqrt{\sigma_{xx}} = \sigma_x$

2.3 Correlation

With the Covariance it is difficult to understand when the relationship between X and Y is strong / weak. We note that

$-\sigma_x\sigma_y \leq \sigma_{xy} \leq \sigma_x\sigma_y$ is quivalent to $-1 \leq \frac{\sigma_{xy}}{\sigma_x\sigma_y} \leq 1$

Correlation between X and Y:

$\rho_{xy} = \frac{\sigma xy}{\sigma_x\sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

- values between -1 and 1
- $\rho_{xy} \approx 0$: there is no dependency between X and Y
- $\rho_{xy} \approx 1(-1)$: there is a strong positive (negative) dependency between X and Y

3 The (simple) linear model

3.1 Linear Trend, the least squares method

We describe the relationship between
`Reaction.Time` and `Age` with a straight line.

$Reaction.Time \approx \beta_0 + \beta_1 Age$
$Y = \beta_0 + \beta_1 X$

Let's draw a line 'in the middle' of the data.

The **least-squares estimator**

We look for the one that passes more 'in the middle', the one that minimizes the sum of the squares of the residues:

$\hat{\beta}_0$ and $\hat{\beta}_1$ such that
$\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ is minimum.

Estimates:

- Angular coefficient: $\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_{xx}} = \rho_{xy}\frac{\sigma_y}{\sigma_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.2064719$
- Intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 10.3013483$
- Response (estimated y): $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- Residuals (from the estimated response): $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$

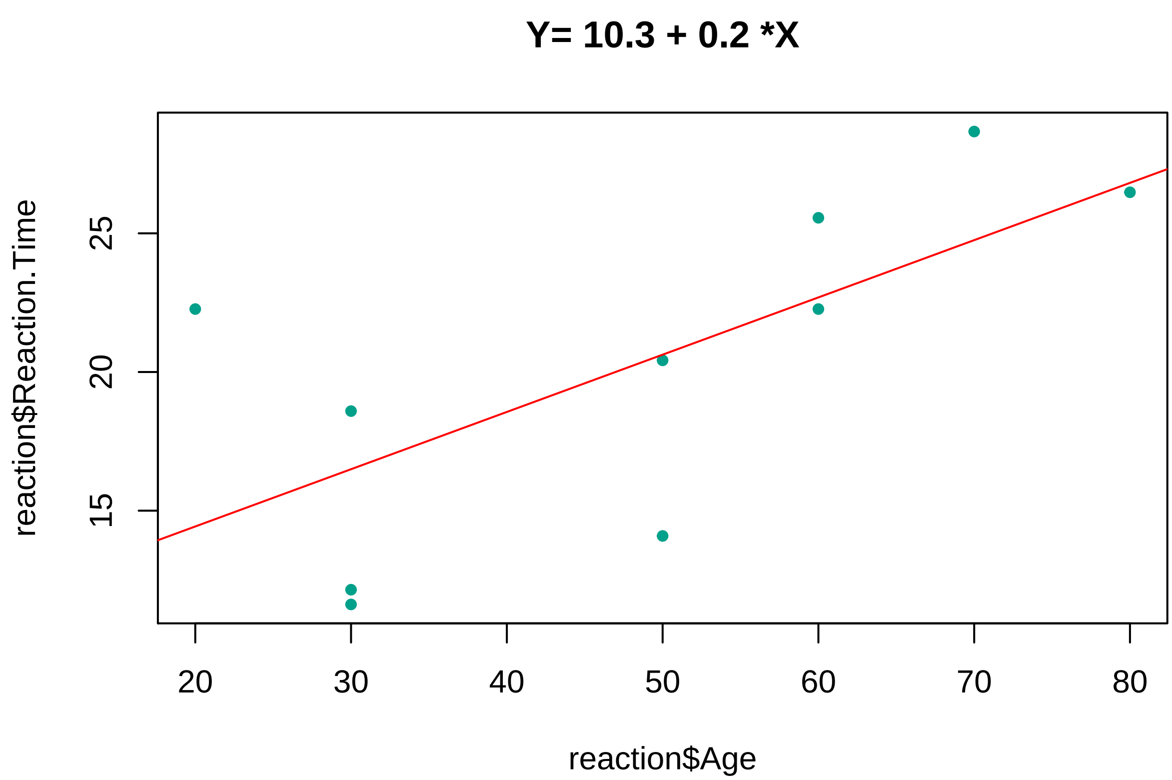
and therefore the least squares are the sum of the squared residuals: $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

A graphical representation:

```
model=lm(Reaction.Time~Age,data=reaction)
coefficients(model)
```

```
## (Intercept)         Age
##  10.3013483   0.2064719
```

```
plot(reaction$Age,reaction$Reaction.Time,pch=20,col=2,cex=1)
coeff=round(coefficients(model),1)
title(paste("Y=",coeff[1],"+",coeff[2],"*X"))
abline(model,col=1)
```



3.2 Interpretation of the coefficients

- β_0 indicates the value of y when $x = 0$ (where the line intersects the ordinate axis).
- β_1 indicates how much y grows as a unit of x grows
  - If $\beta_1 = 0$ there is no relation between x and y.$Y$is constant (horizontal), knowing$x$does not change the estimate of y
  - If $\beta_1 > (<) 0$ the relation between x and y is positive (negative). When X passes from x a $x + 1$ the estimate of Y changes from $\hat{y}$ to $\hat{y} + \hat{\beta}_1$

3.3 The normal (simple) linear model

We assume that the observed values are distributed around true values $\beta_0 + \beta_1 X$ according to a Gaussian law:

$Y =$ linear part $+$ normal error

$Y = \beta_0 + \beta_1 X + \varepsilon$

Assumptions of the linear model

- the $\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$ the relationship between X and the true (mean) Y is linear.
- the **observations** are **independent** each others (knowing the value of the $y_i$observation does not help me to predict the value of y_{i+1}). The random part is ε_i, these are the independent terms.
- $\boldsymbol{\varepsilon_i \sim N(0, \sigma^2),\ \forall i = 1, \ldots, n}$ errors have normal distribution with zero mean and common variance (homoschedasticity: same variance).

3.4 Hypothesis testing

If these assumptions are true,

$\hat{\beta_1} \sim N(\beta_1, \sigma^2 / \sum (x_i - \bar{x})^2)$

We calculate the test statistic:

$t = \frac{\hat{\beta_1}}{std.dev\ \hat{\beta_1}} = \frac{\hat{\beta_1}}{\sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \sum (x_i - \bar{x})^2 / (n-2)}}$

If $H_0 : \beta_1 = 0$, $t \sim t(n-2)$ is true

On `reaction` data and $H_1 : \beta_1 \neq 0$ (bilateral alternative)

```r
model=lm (Reaction.Time ~ Age, data=reaction)
summary (model)
```

```
##
## Call:
## lm(formula = Reaction.Time ~ Age, data = reaction)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.535 -3.364 -0.272  2.676  7.839
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30135    4.04407   2.547   0.0343 *
## Age          0.20647    0.07841   2.633   0.0300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.678 on 8 degrees of freedom
## Multiple R-squared:  0.4643, Adjusted R-squared:  0.3973
## F-statistic: 6.934 on 1 and 8 DF,  p-value: 0.03003
```

Similar result, but much more assumptions!

3.5 The Multiple Linear model

The simple linear model is ‘easily’ extensible to the Multiple Linear Model. Formally we have the same elements, we only expect the linear combination of multiple variables.

$Y =$ linear part $+$ normal error

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p x_p + \varepsilon$

Thus we describe a (hyper) plan of size p.

Assumptions of Multiple linear model

They are the same as the simple linear model

- i) $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$ the relationship between X and Y is truly linear, less than the error termε_i
- ii) the **observations** are among them **independent**
- iii) $\boldsymbol{\varepsilon_i \sim N(0, \sigma^2),\ \forall i = 1, \ldots, n}$

(we will return to the multiple model later)

3.6 Linear regression in R

`> lm (formula, ...)`

where: `formula` specifies the link between the employee and the independent (or predictors)

3.7 Examples of regression model specification

Let y be the dependent variable and x and z two predictors

Regression	Regression in R
$y = \beta_0 + \beta_1 x + \varepsilon$	$lm(y \sim x)$
$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$	$lm(y \sim x + z)$
$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \varepsilon$	$lm(y \sim x + z + x : z)$
$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \varepsilon$	$lm(y \sim x * z)$

For other options on specifying an R model, see: `>? formula`

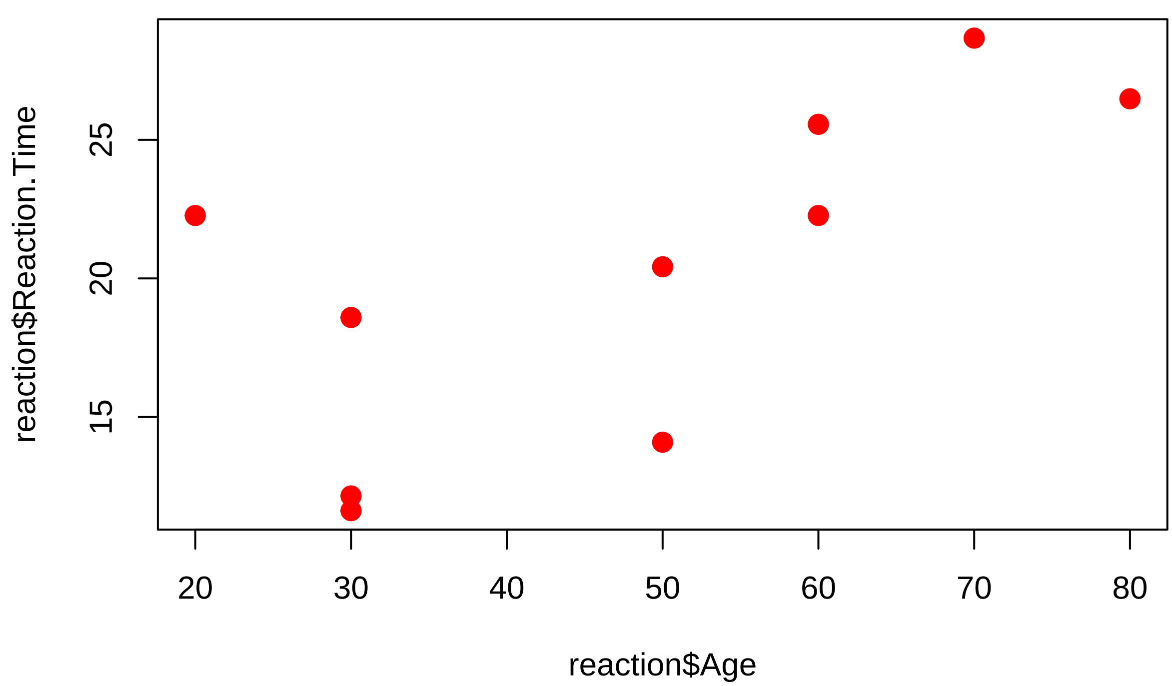
3.8 Basic steps of a regression model

Step	**Code R**	**Libraries**
Model construction	$model = lm(formula)$	stats
Check recruitment	$plot(model)$	stats
Evaluation of parameters	$summary(model)$	stats
Analysis of variance	$anova(model)$	stats
Analysis of variance	$Anova(model, type = “III'')$	car
Viewing effects	see $?effect$	effects
Comparison with other models *	$anova(model, model2)$	stats
Comparison with other models * *	$AIC(model)$;$AIC(model2)$	stats

* comparison between *nested* models based on the$F$test * * model comparison based on the Akaike Information Criterion (AIC) or on the Bayesian Information Criterion (BIC): see also **? AIC**

3.9 Let's go back to our example (simple linear model)

```
plot (reaction$Age, reaction$Reaction.Time, pch = 20, col = 1, cex = 2)
```




```
# to identify observations on the graph with the mouse
# identify (reaction$Age, reaction$Reaction.Time)
```

3.10 Estimate of the model and evaluation of the parameters

```r
model = lm (Reaction.Time ~ Age, data = reaction)
summary (model)
```

```
##
## Call:
## lm(formula = Reaction.Time ~ Age, data = reaction)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.535 -3.364 -0.272  2.676  7.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30135    4.04407   2.547   0.0343 *
## Age          0.20647    0.07841   2.633   0.0300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.678 on 8 degrees of freedom
## Multiple R-squared:  0.4643, Adjusted R-squared:  0.3973
## F-statistic: 6.934 on 1 and 8 DF,  p-value: 0.03003
```

(for now) Note that the test$F$has the same significance as the t test.

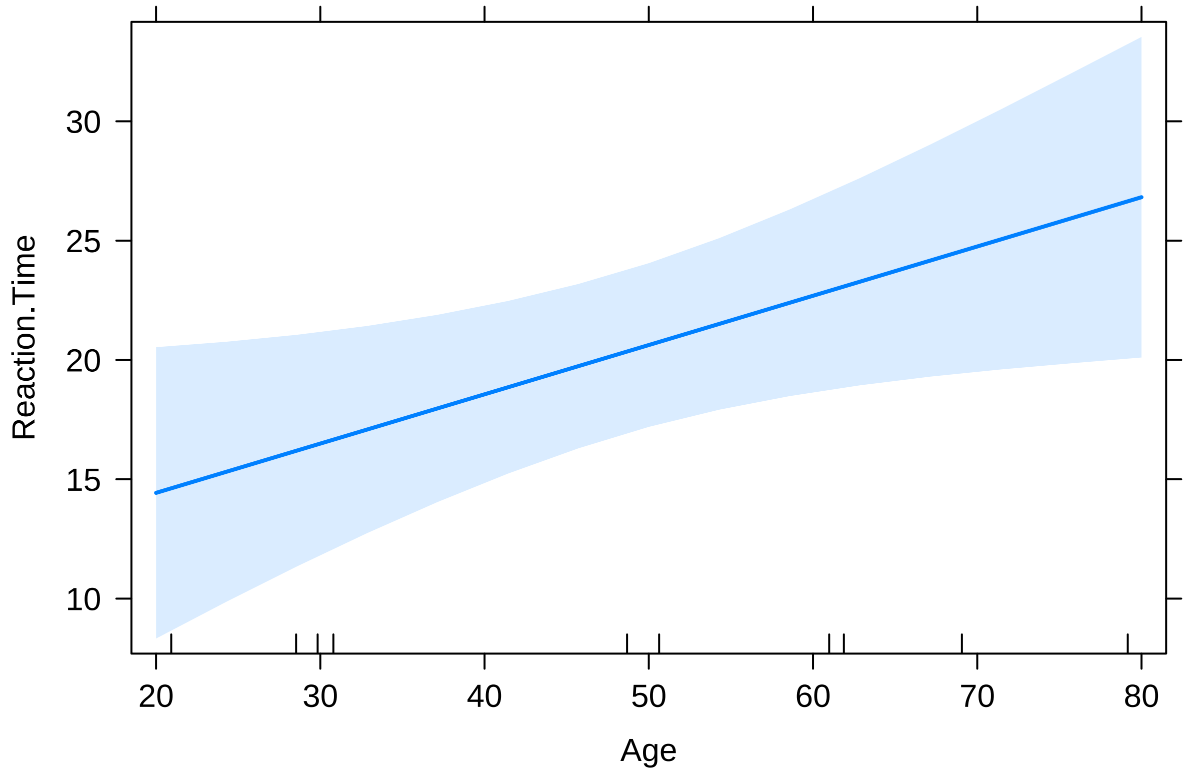
3.11 Graphical representation of the effect of the Age

```r
library (effects) # see:? effect
```

```
## Loading required package: carData
```

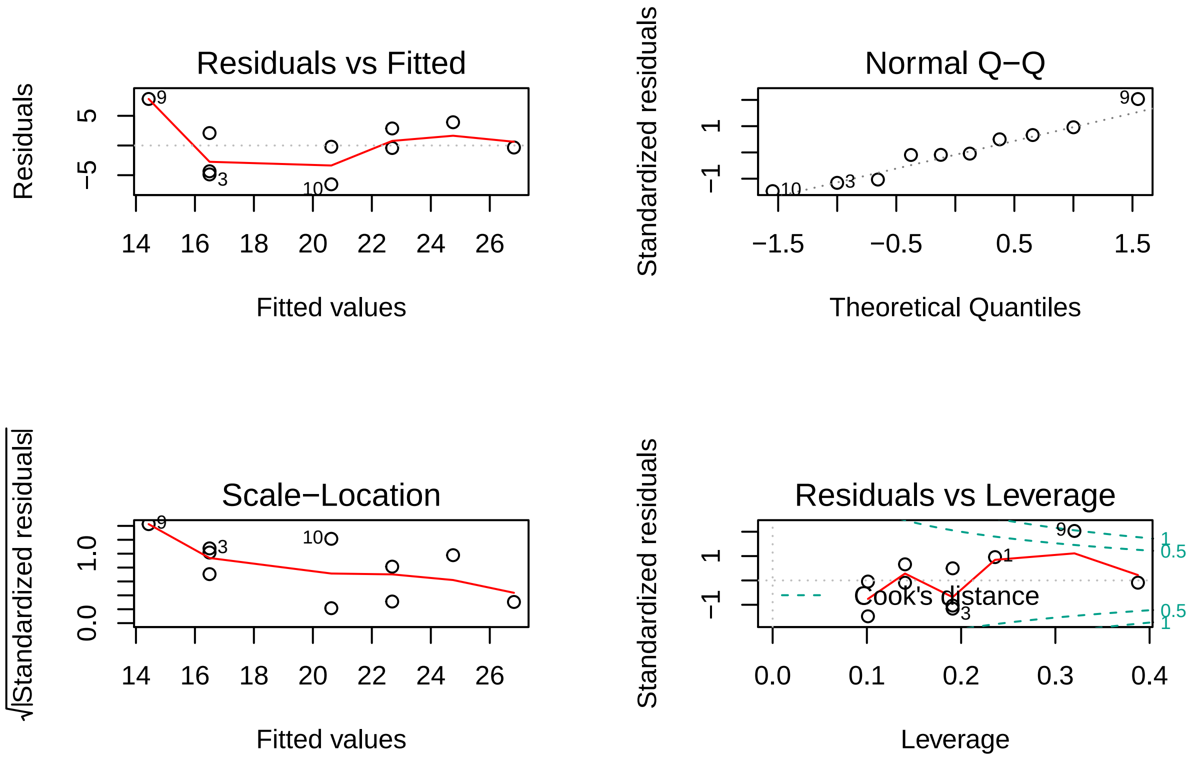
```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
eff <- allEffects (model)
plot (eff, 'Age', ask = F, main = '')
```



3.12 Evaluation of the assumptions on the residuals of the model

```
par (mar = c (6, 5, 4, 2) + 0.1)
par (mfrow = c (2,2))
plot (model) # see also:? plot.lm for bibliographical references
```



- Residual independence?
- Residual conditions?
- Homogeneity variance residues?
- Presence of influential cases?

Please, no test of normality, homoschedasticity etc. (check the error of the first type on the contrary to what you would like).

3.13 Supplement: Looking for influential cases

- In a statistical model an *influential case* is a statistical unit whose observations are strong impact on model parameter estimates
- In regression models, a particularly effective way to identify influential values is to use *Cook's distance (Cook, 1977)*
- Given a statistical unit, Cook's distance is a measure of how much the regression coefficients of the estimated model would change if this unit was omitted
- Greater is Cook's distance, the more the statistical unit helps to determine the parameters of the regression model

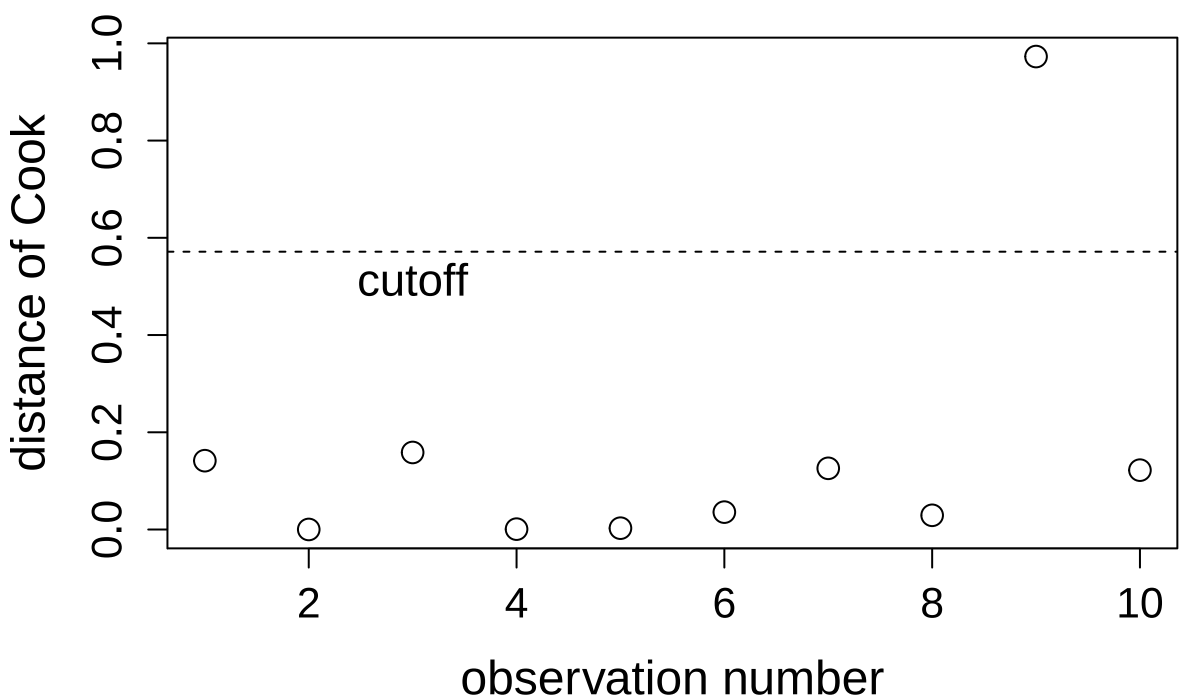
3.14 Identification of influential cases

- In the graph just seen R signals the statistical units with Cook distance values close to 0.5 and to 1, values to be considered as attention thresholds.

- Fox, 2010, proposes a cut-off for Cook's distance that takes into account the number of observations (n) and the number of parameters (k) of the model: $\frac{4}{(n-k-1)}$

3.15 In our case...

```
# calculation and representation of Cook's distance
distances.cook = cooks.distance (model)
plot (distances.cook, xlab = "observation number", ylab = "distance of Cook", cex = 1.5, cex.axis = 1.3
# representation of the cutoff line at the value 4 / (n-k-1)
n = nrow (reaction); k = length (coefficients (model))
cutoff = 4 / (n-k-1)
abline (h= cutoff, lty = 2)
text (3, cutoff * .9, "cutoff", cex = 1.4)
```



3.16 Remark

- Cook's distance is not the only useful indicator for evaluating influential cases. For an overview see R:? Influence.measures
- The identification, evaluation and interpretation of influential cases are fundamental phases of statistical modeling.
- However these aspects are often underestimated in concrete case applications :-(

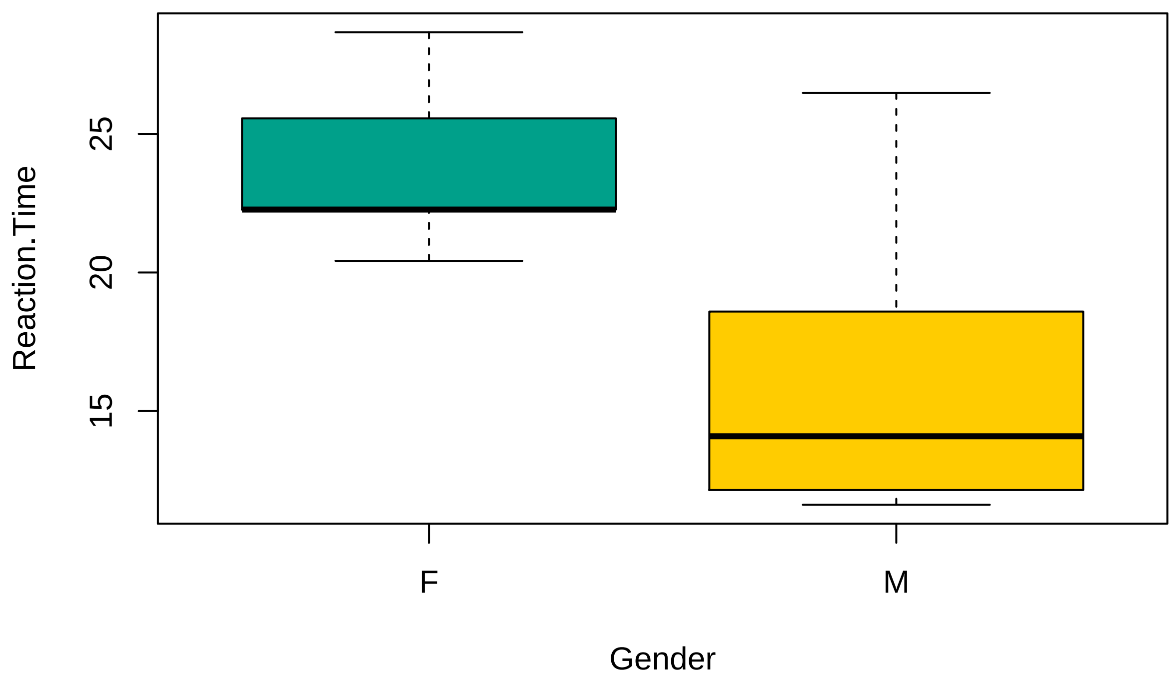
3.16.1 Exercise 1.

Build a regression model by eliminating observation 10. What changes?

4 The Two-independent-samples problem

4.1 The Two-independent-samples problem

```
plot (Reaction.Time ~ Gender, data = reaction, col = 2:3)
```



Is it possible to estimate a model that uses `Gender` as a predictor? How?

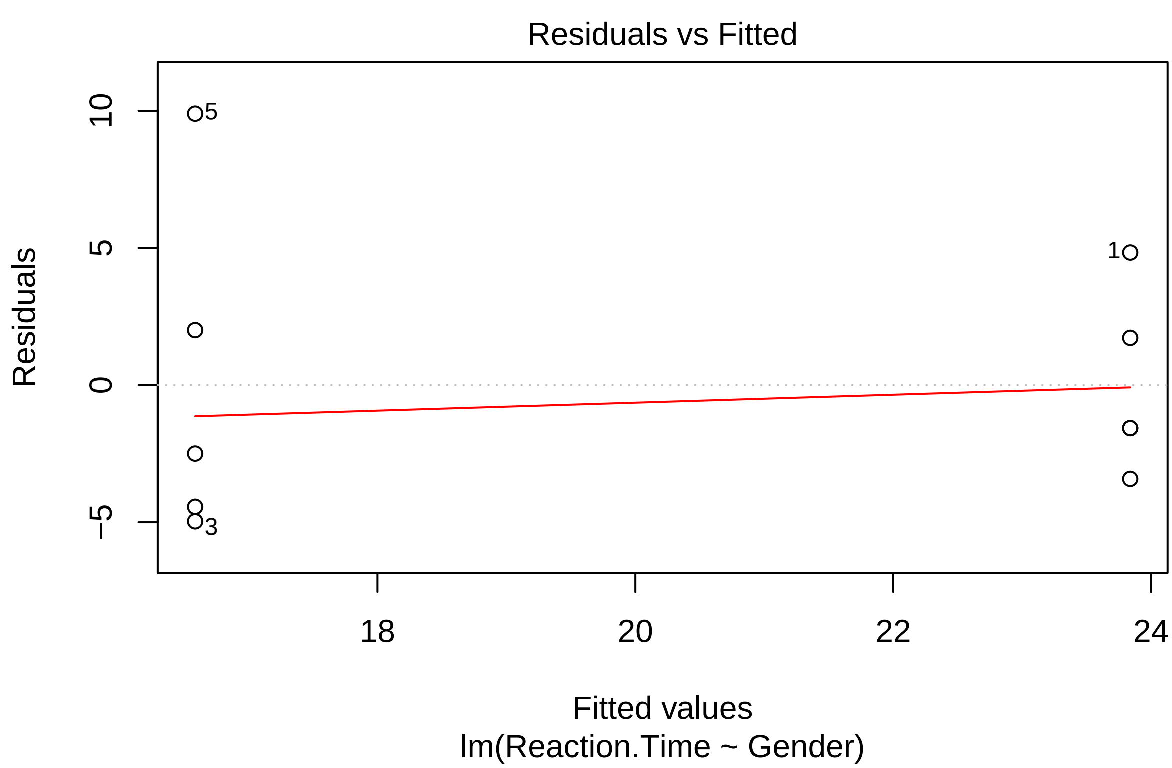
Use `Gender` as if it where a quantitative variable:

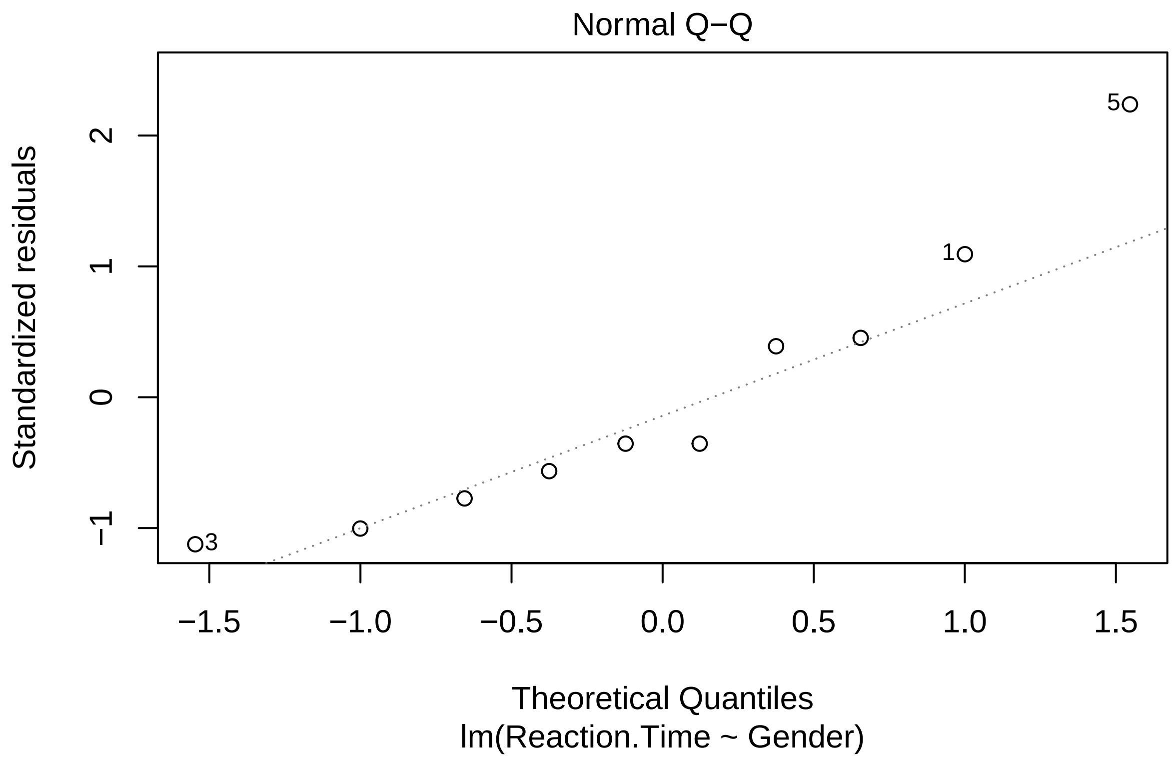
```
modelGender = lm (Reaction.Time ~ Gender, data = reaction)
summary (modelGender)
```

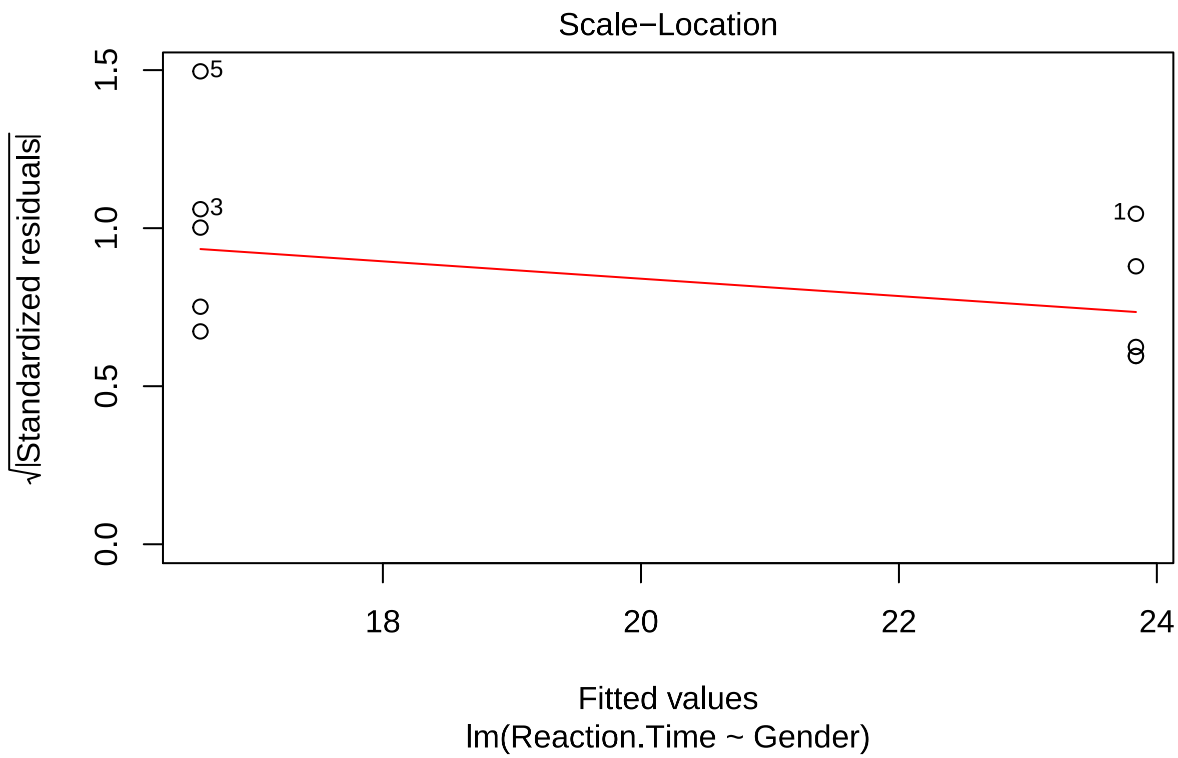
```
##
## Call:
## lm(formula = Reaction.Time ~ Gender, data = reaction)
##
## Residuals:
```

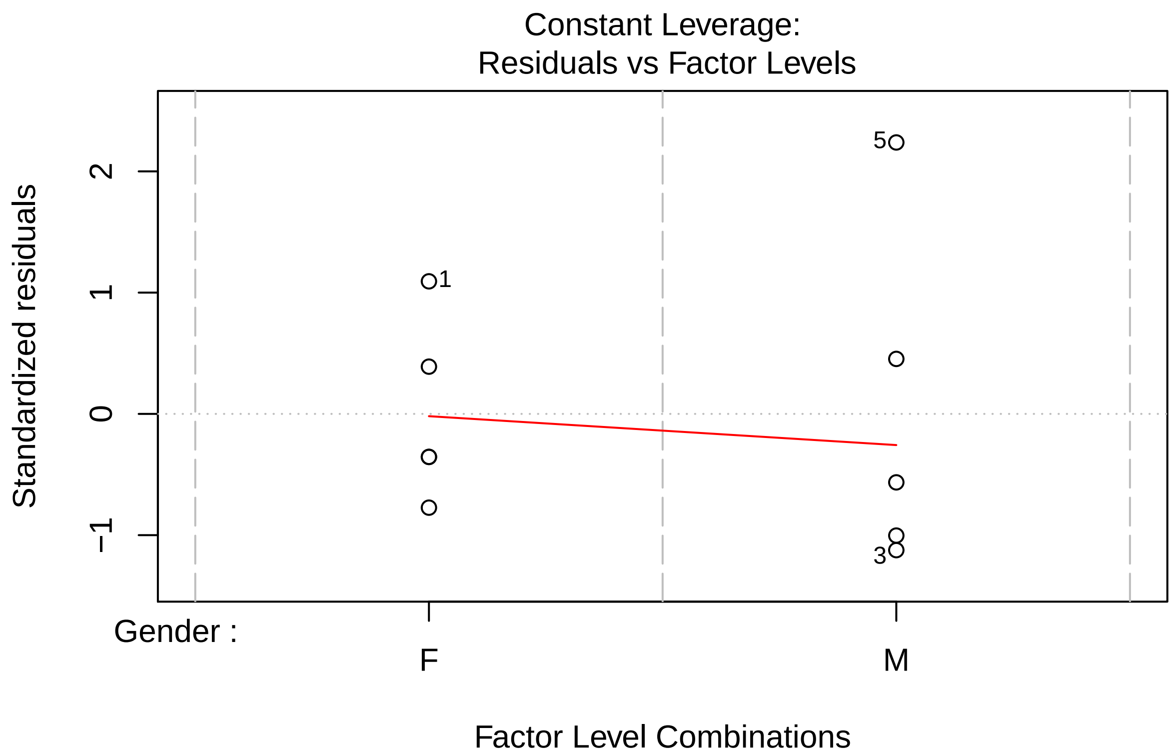
```
##     Min      1Q Median      3Q     Max
## -4.966 -3.188 -1.568  1.933  9.894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.838      2.210   10.79 4.81e-06 ***
## GenderM       -7.252      3.126   -2.32   0.0489 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.942 on 8 degrees of freedom
## Multiple R-squared:  0.4022, Adjusted R-squared:  0.3275
## F-statistic: 5.383 on 1 and 8 DF,  p-value: 0.04891
```

```r
plot (modelGender)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Reaction.Time ~ Gender)
−1
0
1
2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
3
1
5


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
3
1
18
20
22
24
Fitted values
lm(Reaction.Time ~ Gender)



. How do we interpret the coefficients? . What kind of model are we estimating? . What are the differences with my old friend t-test for two independent samples ??

```
by (reaction$Reaction.Time, reaction$Gender, mean)
```

```
## reaction$Gender: F
## [1] 23.838
## --------------------------------------------------------
## reaction$Gender: M
## [1] 16.586
```

```
t.test (Reaction.Time ~ Gender, data = reaction, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Reaction.Time by Gender
## t = 2.3202, df = 8, p-value = 0.04891
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.0443075 14.4596925
## sample estimates:
## mean in group F mean in group M
##          23.838          16.586
```

5 The Multiple linear model

5.1 The Multiple linear model

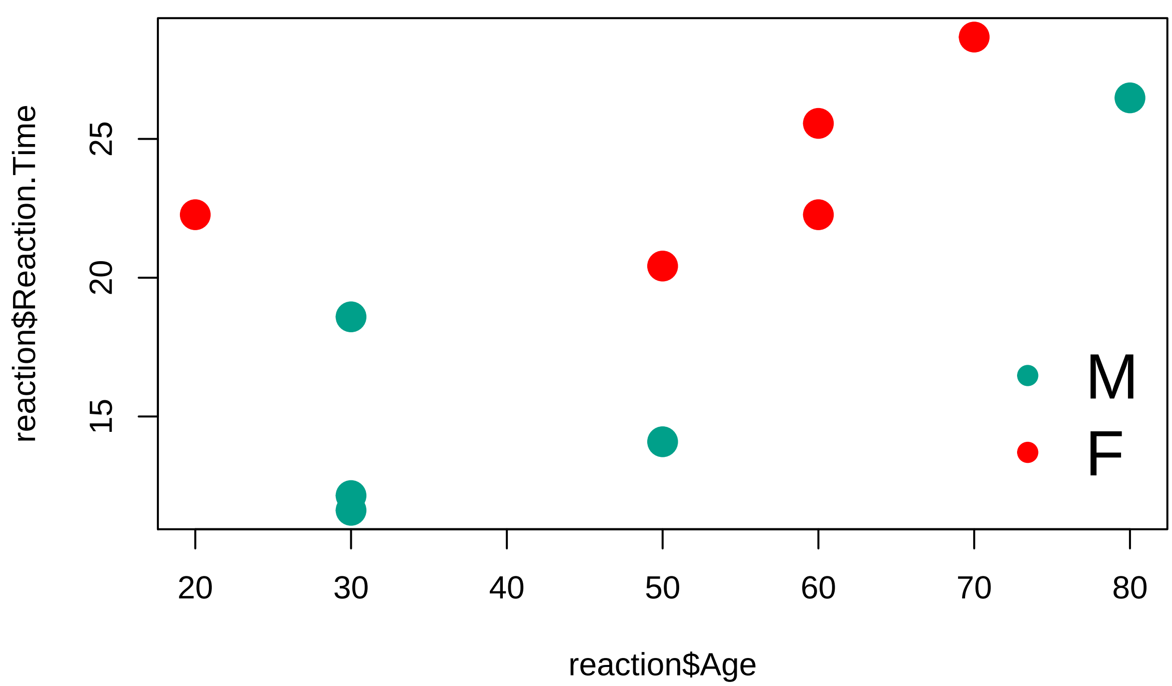
$$Y = \beta_0 + \beta_1 X1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

where is it:

-Y=$Reaction.Time$, height -X_1=Age, shoe size -X_2=$Gender$

5.1.1 Plot the relationship between *Reaction.Time* and *Age* also considering the *Gender*.

```
plot (reaction$Age, reaction$Reaction.Time, col = (reaction$Gender == "M") + 1, pch = 20, cex = 3)
legend ( "bottomright", legend = c ( "M", "F"), pch = 20, cex = 2, col = c (2,1), bty = "n")
```



We know how to estimate a linear model that includes *Reaction.Time* through the *Age*. **EXERCISE:** do it.

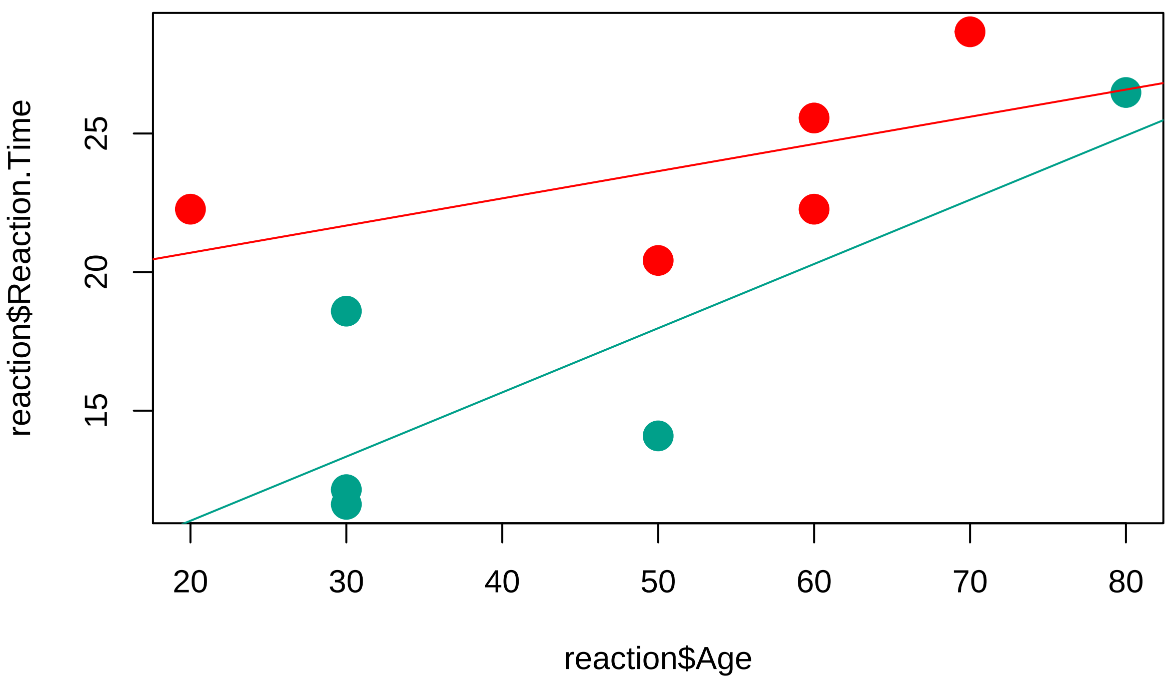
5.2 Multiple linear model

How to estimate a model with *Age*, *Gender* and their interaction?

```r
modelFull = lm (Reaction.Time ~ Age + Gender + Age: Gender, data = reaction)
```

How do we interpret the model?

```r
plot (reaction$Age, reaction$Reaction.Time, col = (reaction$Gender == "M") + 1, pch = 20, cex = 3)
abline (coefficients (modelFull) [1], coefficients (modelFull) [2], col = 1)
abline (coefficients (modelFull) [1] + coefficients (modelFull) [3], coefficients (modelFull) [2] + coe
```



How do we interpret the results of the analysis?

```r
summary (modelFull)
```

```
##
## Call:
## lm(formula = Reaction.Time ~ Age + Gender + Age:Gender, data = reaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8859 -2.1954 -0.1279  1.5675  5.2472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.73568    5.13680   3.647   0.0107 *
## Age          0.09812    0.09378   1.046   0.3358
```

```
## GenderM     -12.34255    6.48970  -1.902   0.1059
## Age:GenderM   0.13353    0.12480   1.070   0.3258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.608 on 6 degrees of freedom
## Multiple R-squared:  0.7611, Adjusted R-squared:  0.6416
## F-statistic:  6.37 on 3 and 6 DF,  p-value: 0.02703
```

The F test (shown below in the table) tests the hypothesis: $H_0: \ \beta_1 = \ldots = \beta_p = 0$ (all equal to 0) versus $H_0: \text{ At least one } \beta_i \neq 0$(at least one other than 0)

In this case we have reason to believe that there is at least one useful predictor between Gender, Age and their interaction ($p < .05$).

The coefficients are estimated and tested net of the effect of the other variables …

5.2.1 Correlation between predictors

In the multiple regression models we lose the relationship between correlation andR^2 (among other things there are p possible correlations with Y).

The estimation of the coefficients is done in a joint manner, therefore affected by the correlation between the predictors X

```
cor (reaction$Age, reaction$Gender == "M")
```

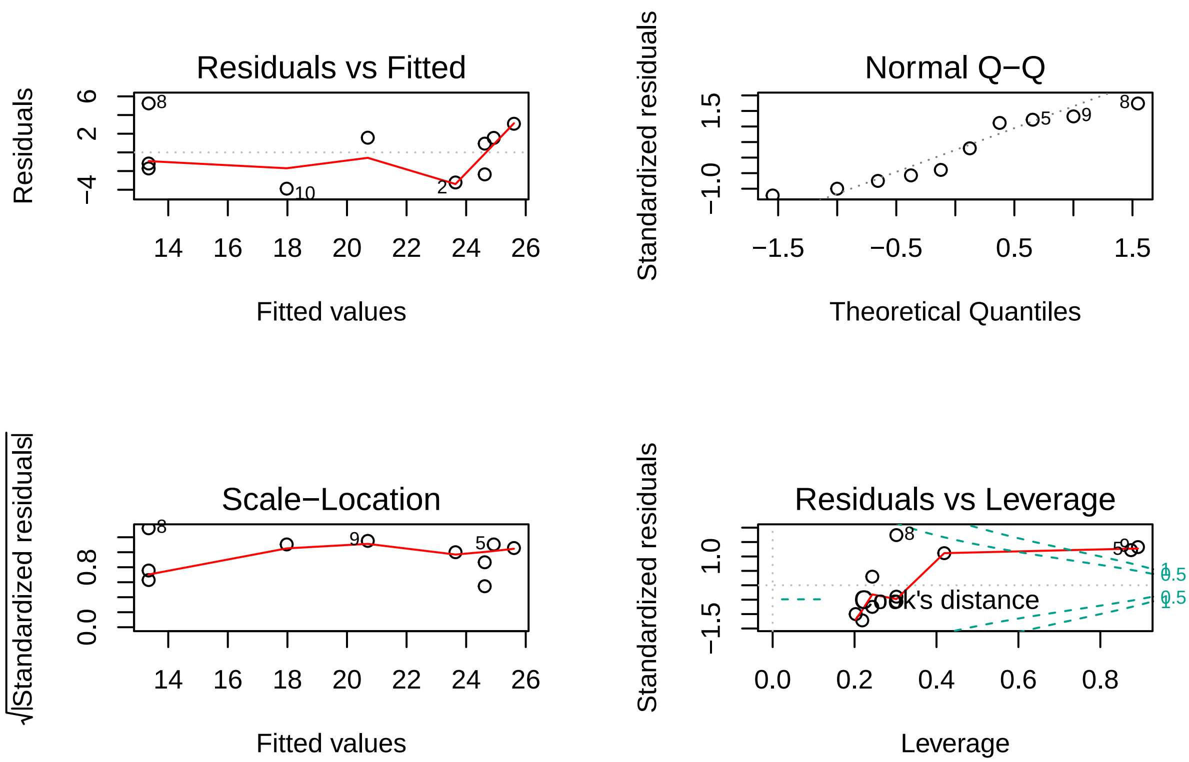
```
## [1] -0.2119996
```

it is very high, this will bring instability (greater variance) in the estimates that will be less precise (and therefore higher p-values, wider confidence intervals).

This is the main reason why it is useful to have experiments with orthogonal factorial plans (not discussed today)

5.2.2 Residual Analysis

```
par (mar = c (6, 5, 4, 2) + 0.1)
par (mfrow = c (2,2))
plot (modelFull) # see also:? plot.lm for bibliographical references
```



5.3 Analysis of variance

The Deviance Explained and (andR^2) increases - does not decrease - with each addition of variables (+ varibili = + flexibility = better fit).

REMARK: this mean that we are considering **nested models**

for example:

```r
summary (modelFull)
```

```
##
## Call:
## lm(formula = Reaction.Time ~ Age + Gender + Age:Gender, data = reaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8859 -2.1954 -0.1279  1.5675  5.2472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.73568    5.13680   3.647   0.0107 *
## Age          0.09812    0.09378   1.046   0.3358
## GenderM    -12.34255    6.48970  -1.902   0.1059
## Age:GenderM  0.13353    0.12480   1.070   0.3258
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.608 on 6 degrees of freedom
## Multiple R-squared:  0.7611, Adjusted R-squared:  0.6416
## F-statistic:  6.37 on 3 and 6 DF,  p-value: 0.02703
```

```r
modelAgeGen = lm (Reaction.Time ~ Age + Gender, data = reaction)
summary (modelAgeGen)
```

```
##
## Call:
## lm(formula = Reaction.Time ~ Age + Gender, data = reaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5372 -2.8513 -0.8364  3.1623  4.4334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.81447    3.63652   4.074  0.00473 **
## Age          0.17353    0.06251   2.776  0.02746 *
## GenderM     -5.86376    2.35899  -2.486  0.04186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.645 on 7 degrees of freedom
## Multiple R-squared:  0.7155, Adjusted R-squared:  0.6342
## F-statistic: 8.801 on 2 and 7 DF,  p-value: 0.01229
```

```r
modelAge = lm (Reaction.Time ~ Age, data = reaction)
summary (modelAge)
```

```
##
## Call:
## lm(formula = Reaction.Time ~ Age, data = reaction)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.535 -3.364 -0.272  2.676  7.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.30135    4.04407   2.547   0.0343 *
## Age          0.20647    0.07841   2.633   0.0300 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.678 on 8 degrees of freedom
## Multiple R-squared:  0.4643, Adjusted R-squared:  0.3973
## F-statistic: 6.934 on 1 and 8 DF,  p-value: 0.03003
```

From the analysis it seems that the interaction and the Gender are not predictive. We test this hypothesis through a comparison of nested models

```r
anova (modelAgeGen, modelFull)
```

```
## Analysis of Variance Table
##
## Model 1: Reaction.Time ~ Age + Gender
## Model 2: Reaction.Time ~ Age + Gender + Age:Gender
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      7 93.008
## 2      6 78.105  1    14.903 1.1448 0.3258
```

Among the multiple models with or without interaction there is no significant difference in terms of the explained variance.

With ANOVA test we make the following question: "Does the exclusion of predictor X decreases the predictability of the response?". This evaluation is not only based on the reduction of Residual Standard Error (i.e. decrease of Multiple R-squared), but also the reducted flexiblity of the model (i.e. the DF spent to model the tested variable X).

As index, the Adjusted R-squared is a more "honest" index of explained variance then the Multiple R-squared.

Excluding the *Gender* variable instead does not seem like a good idea:

```r
anova (modelAge, modelAgeGen)
```

```
## Analysis of Variance Table
##
## Model 1: Reaction.Time ~ Age
## Model 2: Reaction.Time ~ Age + Gender
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1      8 175.104
## 2      7  93.008  1    82.096 6.1788 0.04186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

… and not even removing *Age*:

```r
anova (modelGender, modelAgeGen)
```

```
## Analysis of Variance Table
##
## Model 1: Reaction.Time ~ Gender
## Model 2: Reaction.Time ~ Age + Gender
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1      8 195.390
## 2      7  93.008  1    102.38 7.7056 0.02746 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The best (most parsimonious) model is the one with only*Age*and*Gender*but without interaction.

5.4 Model selection via AIC and BIC

These are methods that penalize models with many predictors.

We compare the BIC (Bayesian Information Criterion) or the AIC (Akaike Information Criterion) of the models. The idea: the lower the BIC and the better the model

```
n = nrow (reaction)
(BIC1 = AIC (modelFull, k = log (n)))
```

```
## [1] 60.44635
```

```
(BIC2 = AIC (modelAgeGen, k = log (n)))
```

```
## [1] 59.89008
```

```
(BIC3 = AIC (modelAge, k = log (n)))
```

```
## [1] 63.91446
```

```
(BICGender = AIC (modelGender, k = log (n)))
```

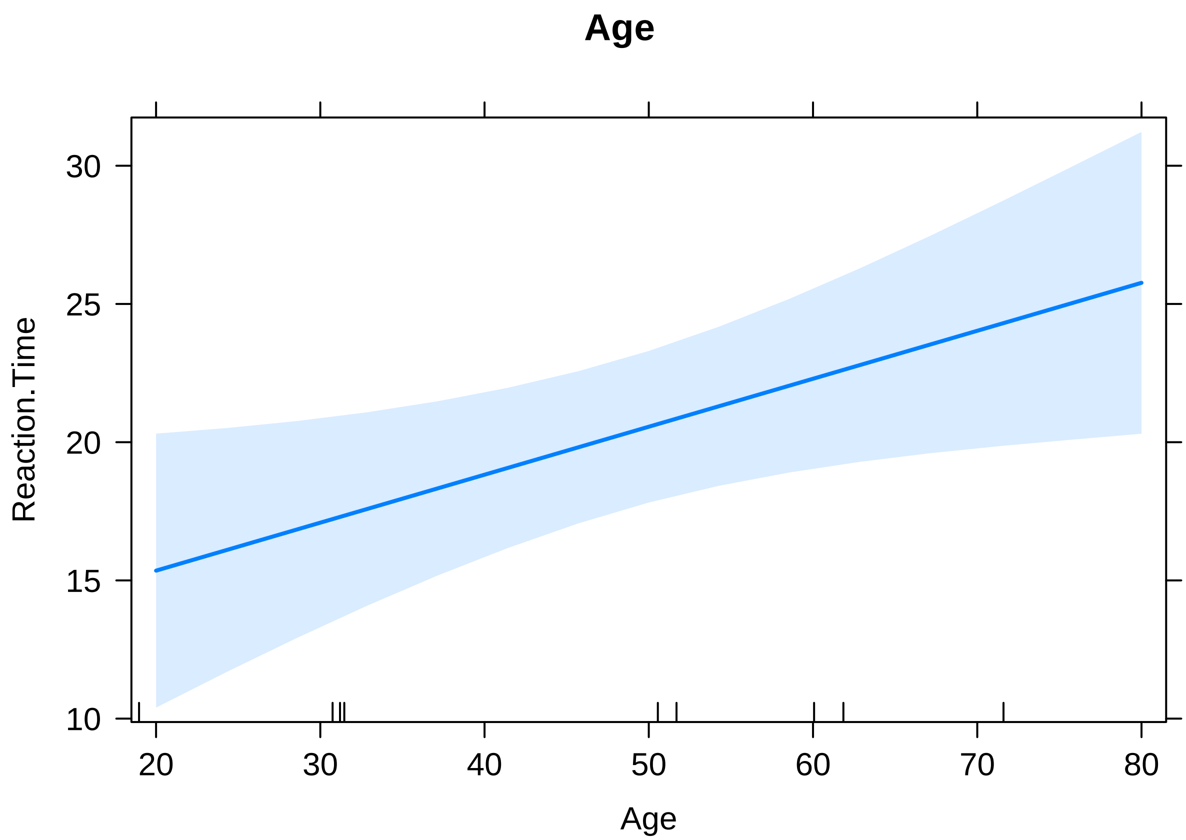
```
## [1] 65.01065
```

(Also in this case) The model with `Age + Gender` seems to be the best.

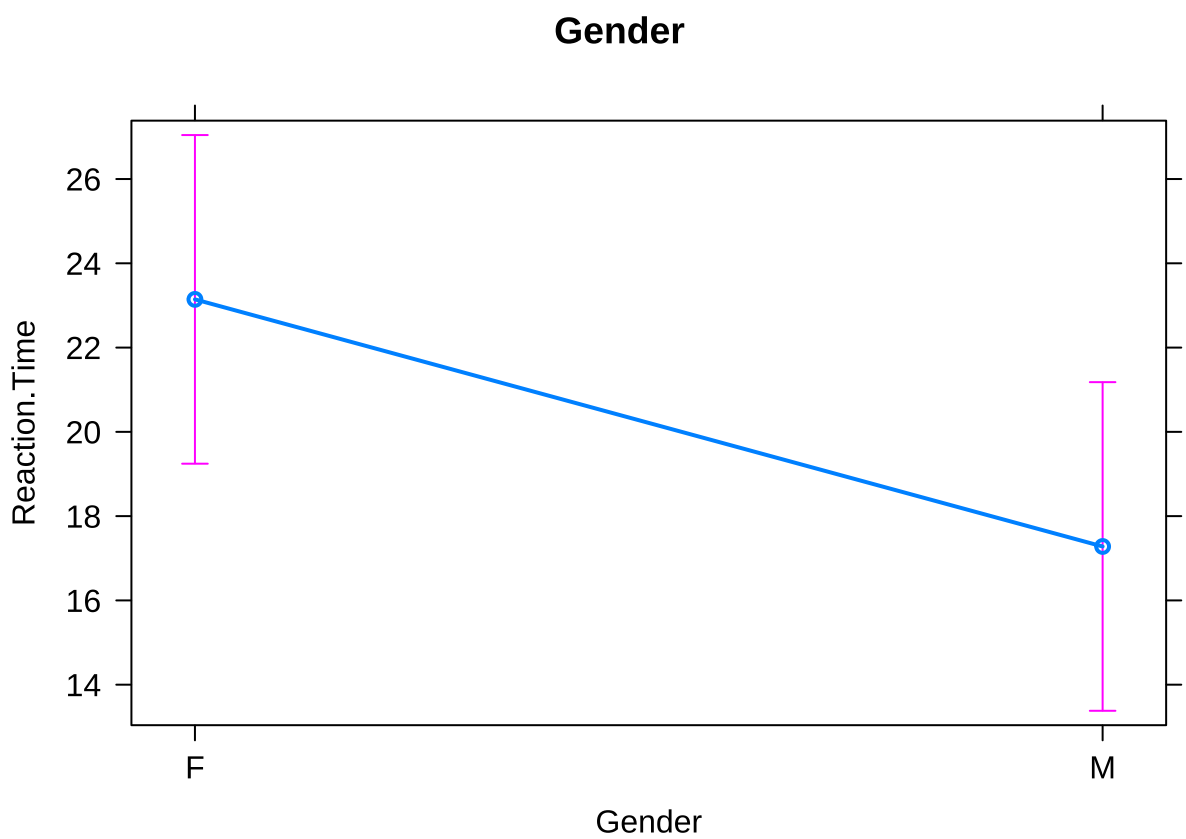
```
summary (modelAgeGen)
```

```
##
## Call:
## lm(formula = Reaction.Time ~ Age + Gender, data = reaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5372 -2.8513 -0.8364  3.1623  4.4334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.81447    3.63652   4.074  0.00473 **
## Age          0.17353    0.06251   2.776  0.02746 *
## GenderM     -5.86376    2.35899  -2.486  0.04186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.645 on 7 degrees of freedom
## Multiple R-squared:  0.7155, Adjusted R-squared:  0.6342
## F-statistic: 8.801 on 2 and 7 DF,  p-value: 0.01229
```

```r
eff <- allEffects(modelAgeGen)
par(mfrow=c(1,2))
plot(eff,'Age',ask=F,main='Age')
```




```r
plot(eff,'Gender',ask=F,main='Gender')
```




```
par(mfrow=c(1,1))
```