Part 1 - Linear model

Cathy Maugis-Rabusseau and Pierre Neuvial

Institut de Mathématiques de Toulouse

2021-06-25

Section 1

Introduction

Example

- For 100 individuals, we have their height, weight, age and sex (75 men and 25 women). We also know whether they are smokers or not; whether they snore at night or not.

```
  age weight height sex snore tobacco
1  47     71    158   H     N       O
2  56     58    164   H     O       N
3  46    116    208   H     N       O
4  70     96    186   H     N       O
5  51     91    195   H     O       O
6  46     88    188   F     N       N
```

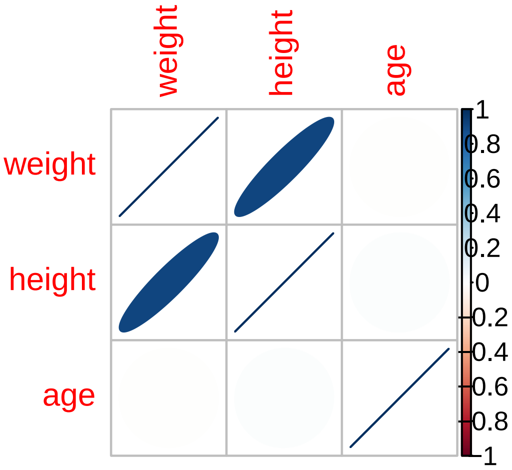
```r
summary(don)
```

```
      age             weight           height       sex     snore    tobacco
 Min.   :23.00   Min.   : 42.00   Min.   :158.0   F:25   N:65   N:36
 1st Qu.:43.00   1st Qu.: 75.50   1st Qu.:166.0   H:75   O:35   O:64
 Median :52.00   Median : 92.00   Median :186.0
 Mean   :52.27   Mean   : 88.83   Mean   :181.1
 3rd Qu.:62.25   3rd Qu.:104.25   3rd Qu.:194.0
 Max.   :74.00   Max.   :120.00   Max.   :208.0
```

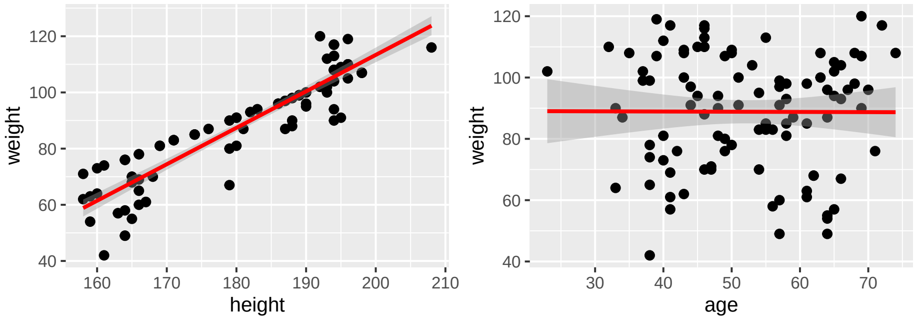
- 3 quantitative variables and 3 qualitative variables

Explain weight ∼ height / age (linear regression)

- Correlation between the quantitative variables:



Explain weight ∼ height / age (linear regression)



Explain weight $\sim$ height (linear regression)

- Model:

$$weight_i = a + b \times height_i + \varepsilon_i, \; i = 1, \cdots, 100$$

  where ε_i is the noise for the i-th observation

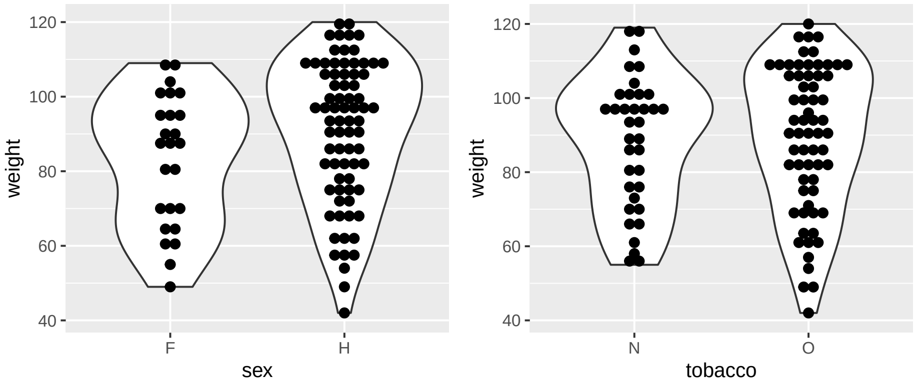
- Assumptions: $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d $\mathcal{N}(0, \sigma^2)$
  (Gaussian error with the same unknown variance)

- Matricial writing:

$$\underbrace{\begin{pmatrix} weight_1 \\ \vdots \\ weight_{100} \end{pmatrix}}_{weight} = \underbrace{\begin{pmatrix} 1 & height_1 \\ \vdots & \vdots \\ 1 & height_{100} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{100} \end{pmatrix}}_{\varepsilon}$$

$$\Leftrightarrow weight = X\theta + \varepsilon, \; \varepsilon \sim \mathcal{N}_n(O_n, \sigma^2 I_n)$$

Explain weight ∼ sex / tobacco (Anova)



Explain weight $\sim$ sex (Anova)

- One-way ANOVA
  - Model per observation:

$$weight_i = \mu_1 \mathbb{1}_{sex_i=F} + \mu_2 \mathbb{1}_{sex_i=H} + \varepsilon_i \text{ where } \varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$$

  - Matricial writing:

$$\underbrace{\begin{pmatrix} weight_{11} \\ \vdots \\ weight_{1n_1} \\ weight_{21} \\ \vdots \\ weight_{2n_2} \end{pmatrix}}_{weight} = \underbrace{\begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}}_{\theta} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}}_{\varepsilon},$$

where $weight_{i,j}$ = weight of the j-th individual with sex $i = F$ or H, $j \in \{1, \ldots, n_i\}$.

Explain weight $\sim$ sex and tobacco (Anova)

- Two-way ANOVA
  - Joint effect of sex and tobacco on weight.
  - Model:

$$weight_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

  where $weight_{ijk}$ = weight of the k-th individual with $sex = i \in \{H, F\}$ and $tobacco = j \in \{0, N\}$, $k \in \{1, \dots, n_{ij}\}$.

- This model can also be written matricially

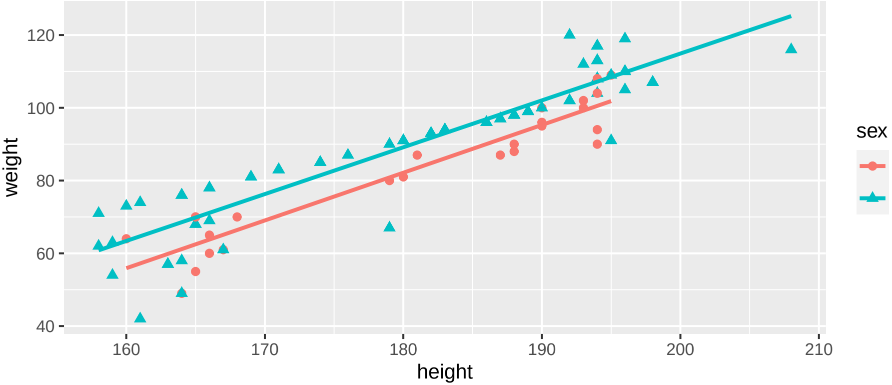
$$weight = X\theta + \varepsilon,\ \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$$

Explain weight $\sim$ sex and height (ANCOVA)

- Model:

$$weight_{ij} = a_i + b_i \; height_{ij} + \varepsilon_{ij}, \;\; i \in \{H, F\} \text{ and } j = 1, \cdots, n_i$$

where $weight_{ij}$ = weight of the j-th individual with sex i,
$\varepsilon_{ij} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$.



Section 2

Linear regression

Notation

- Let Y be a **quantitative** response variable
- Let p **quantitative** explanatory variables $X^{(1)}, \dots, X^{(p)}$
- Data : the observation of a n-sample:

$$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ and } \forall i = 1, \dots, n,\ X_i = (X_i^{(1)}, \dots, X_i^{(p)})$$

- For simplicity, some concepts will be introduced in the linear regression case but they can be extended to the ANOVA and ANCOVA.

Example

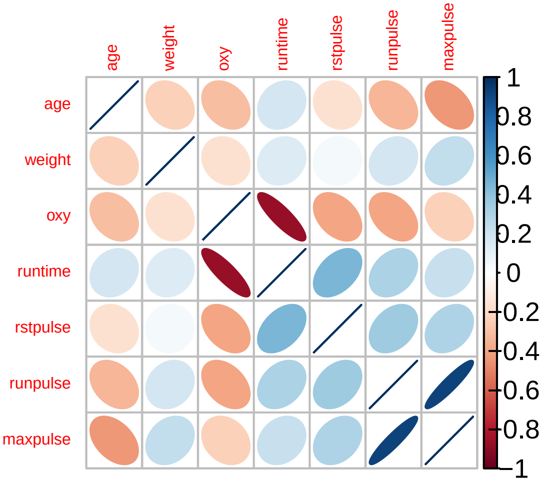
- Data collected for 31 persons during aerobic sessions
- 7 variables:
    - age (a): age
    - weight (w): weight
    - oxy (oxy): oxygen consumption
    - runtime (run): time of effort
    - rstpulse (rst): heart rate measurement 1
    - runpulse (rp): heart rate measurement 2
    - maxpulse (maxp): heart rate measurement 3

Example

We want to explain the consumption of oxygen (response variable Y=oxy) according to the other quantitative variables ($p = 6$).

```
      age            weight           oxy            runtime     
 Min.   :38.00   Min.   :59.08   Min.   :37.39   Min.   : 8.17  
 1st Qu.:44.00   1st Qu.:73.20   1st Qu.:44.96   1st Qu.: 9.78  
 Median :48.00   Median :77.45   Median :46.77   Median :10.47  
 Mean   :47.68   Mean   :77.44   Mean   :47.38   Mean   :10.59  
 3rd Qu.:51.00   3rd Qu.:82.33   3rd Qu.:50.13   3rd Qu.:11.27  
 Max.   :57.00   Max.   :91.63   Max.   :60.05   Max.   :14.03  
    rstpulse        runpulse        maxpulse    
 Min.   :40.00   Min.   :146.0   Min.   :155.0  
 1st Qu.:48.00   1st Qu.:163.0   1st Qu.:168.0  
 Median :52.00   Median :170.0   Median :172.0  
 Mean   :53.45   Mean   :169.6   Mean   :173.8  
 3rd Qu.:58.50   3rd Qu.:176.0   3rd Qu.:180.0  
 Max.   :70.00   Max.   :186.0   Max.   :192.0  
```

Example


age
weight
oxy
runtime
rstpulse
runpulse
maxpulse
age
weight
oxy
runtime
rstpulse
runpulse
maxpulse
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Definition of a linear model

- $\forall i = 1, \ldots, n$,

$$Y_i = \underbrace{\theta_0 + \theta_1 X_i^{(1)} + \ldots + \theta_p X_i^{(p)}}_{(*)} + \underbrace{\varepsilon_i}_{noise} \tag{1}$$

- (*) = average response = **linear combination** of explanatory variables
- Assumptions:
  - $\varepsilon_1, \ldots, \varepsilon_n$ independent and identically distributed (i.i.d)
    $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$
  - Mainly Gaussian errors: $\varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$
- $\theta = (\theta_0, \ldots, \theta_p)'$ and σ^2 are **unknown** parameters

Definition of a linear model

- $Y_i = \theta_0 + \theta_1 X_i^{(1)} + \ldots + \theta_p X_i^{(p)} + \varepsilon_i, \forall i = 1, \ldots, n$
- Matricial writing:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^{(1)} & \ldots & X_1^{(j)} & \ldots & X_1^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_i^{(1)} & \ldots & X_i^{(j)} & \ldots & X_i^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{(1)} & \ldots & X_n^{(j)} & \ldots & X_n^{(p)} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\Leftrightarrow Y = X\theta + \varepsilon$ where $Y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times k}$ with $k = p + 1, \theta \in \mathbb{R}^k$

Least square estimation for θ

- Linear model :

$$Y = X\theta + \varepsilon \text{ with } \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$$

where θ and σ^2 are **unknown** parameters.

- Least square estimation: we minimise

$$\|Y - X\theta\|^2 = \sum_{i=1}^{n} (Y_i - (X\theta)_i)^2$$

  - If $X'X$ invertible: $\hat{\theta} = (X'X)^{-1}X'Y$ (unique)
  - If $X'X$ is not invertible (in particular if $p > n$): θ is not uniquely defined (model not identifiable) $\longrightarrow$ Additional constraints, Variable selection, ...

- In the sequel, we assume that the model is regular (X'X invertible $\Leftrightarrow rank(X) = k$)

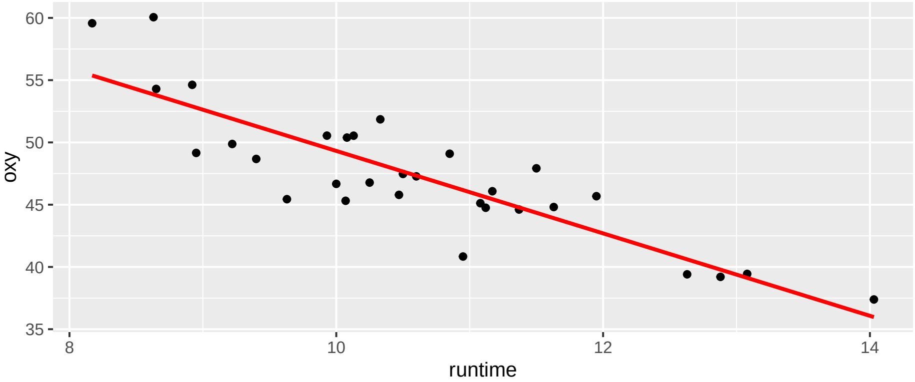
Example (simple linear regression)

```r
reg1 = lm(oxy~runtime,data=fitness)
summary(reg1)
```

```
Call:
lm(formula = oxy ~ runtime, data = fitness)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3352 -1.8424 -0.0569  1.5342  6.2033

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  82.4218     3.8553  21.379  < 2e-16 ***
runtime      -3.3106     0.3612  -9.166 4.59e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.745 on 29 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7345
F-statistic: 84.01 on 1 and 29 DF,  p-value: 4.585e-10
```

Example (simple linear regression)

```
ggplot(fitness,aes(x=runtime,y=oxy))+
  geom_point()+
  geom_smooth(method=lm,se=FALSE,col="red")
```



$$\begin{cases} \hat{\theta}_1 = cov(oxy, runtime)/var(runtime) \\ \hat{\theta}_0 = \overline{oxy} - \hat{\theta}_1 \overline{runtime} \end{cases}$$

Example (multiple linear regression)

```r
regmulti=lm(oxy~.,data=fitness)
summary(regmulti)
```

```
Call:
lm(formula = oxy ~ ., data = fitness)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.4026 -0.8991  0.0706  1.0496  5.3847 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.93448   12.40326   8.299 1.64e-08 ***
age          -0.22697    0.09984  -2.273  0.03224 *  
weight       -0.07418    0.05459  -1.359  0.18687    
runtime      -2.62865    0.38456  -6.835 4.54e-07 ***
rstpulse     -0.02153    0.06605  -0.326  0.74725    
runpulse     -0.36963    0.11985  -3.084  0.00508 ** 
maxpulse      0.30322    0.13650   2.221  0.03601 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.317 on 24 degrees of freedom
Multiple R-squared:  0.8487,    Adjusted R-squared:  0.8108 
F-statistic: 22.43 on 6 and 24 DF,  p-value: 9.715e-09
```

Predicted values and residuals

- **Predicted values** of Y: $\widehat{Y} = X\widehat{\theta}$

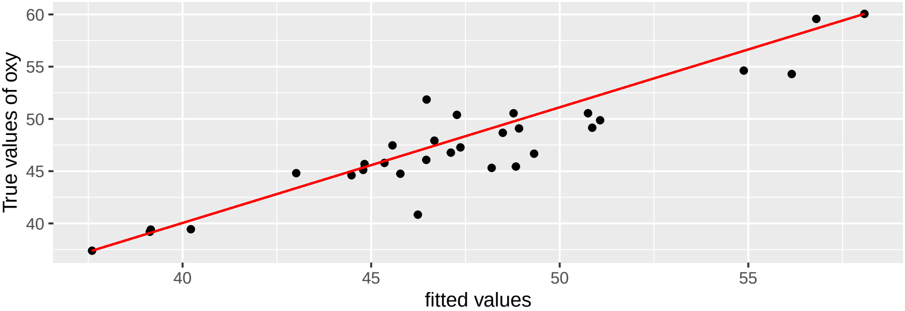
$\forall i \in \{1, \dots, n\},\ \widehat{Y}_i = (X\widehat{\theta})_i = \widehat{\theta}_0 + \widehat{\theta}_1 X_i^{(1)} + \dots + \widehat{\theta}_p X_i^{(p)}$

$=$ projection of Y onto the subspace generated by the columns of X $(Vect(X))$

- **Residuals**: $\widehat{\varepsilon} = Y - \widehat{Y}$ i.e $\forall i,\ \widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$

$=$ the orthogonal projection of Y onto the subspace $Vect(X)^{\perp}$

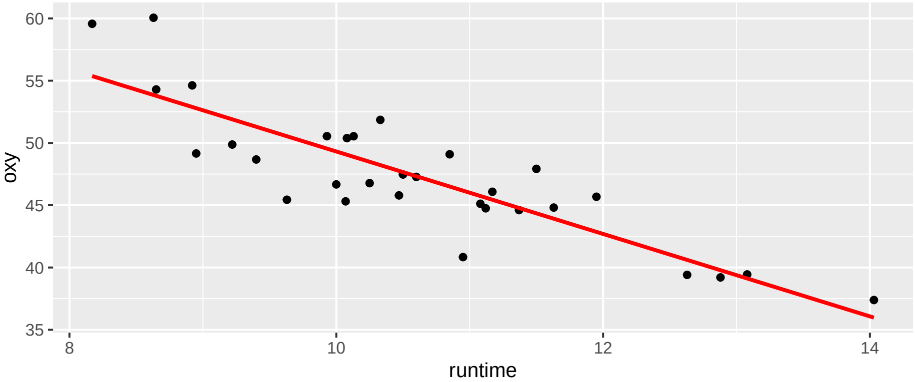
Example



Estimation of the variance σ^2

- σ^2 is the common variance of the errors ε_i
- Unbiased estimator:

$$\widehat{\sigma^2} = \frac{\|\widehat{\varepsilon}\|^2}{n-k} = \frac{\|Y - \widehat{Y}\|^2}{n-k} = \frac{\|Y - X\widehat{\theta}\|^2}{n-k} = \frac{SSR(\widehat{\theta})}{n-k}$$



Properties of the estimators

If $Y = X\theta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$ and $X'X$ invertible,

- $\widehat{\theta} \sim \mathcal{N}_{p+1}(\theta, \sigma^2(X'X)^{-1})$ thus $\hat{\theta}_j \sim \mathcal{N}(\theta_j, \sigma^2[(X'X)^{-1}]_{j,j})$

- $\dfrac{(n-k)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-k)$

- $\widehat{\theta}$ and $\widehat{\sigma^2}$ are independent

With these results, we may build a confidence interval for each θ_j, a test for the significance of a variable $X^{(j)}$, ...

Confidence interval for θ_j

- Based on a Student statistics:

$$\frac{\widehat{\theta}_j - \theta_j}{\sqrt{\widehat{\sigma^2}[(X'X)^{-1}]_{jj}}} \sim \mathcal{T}(n-k).$$

- Expression of the confidence interval for θ_j:

$$IC_{1-\alpha}(\theta_j) = \left[\ \widehat{\theta}_j \pm t_{1-\alpha/2,n-k}\ \sqrt{\widehat{\sigma}^2[(X'X)^{-1}]_{jj}}\ \right] = \left[\widehat{\theta}_j \pm t_{1-\alpha/2,n-k}\ se_j\right]$$

where $t_{1-\alpha/2,n-k}$ is the $1-\alpha/2$ quantile of Student distribution $\mathcal{T}(n-k)$.

```
confint(regmulti,level=0.95)
```

```
                  2.5 %       97.5 %
(Intercept) 77.33541293 128.53354604
age         -0.43302821  -0.02091938
weight      -0.18685216   0.03849733
runtime     -3.42235018  -1.83495545
rstpulse    -0.15786297   0.11479569
runpulse    -0.61699207  -0.12226345
maxpulse     0.02150491   0.58492935
```

Test for the nullity of θ_j

- Test for the significance of the variable $X^{(j)}$
- $\mathcal{H}_0 : \theta_j = 0$ vs $\mathcal{H}_1 : \theta_j \neq 0$
- We reject $\mathcal{H}_0$ at level α if $|T^{(j)}| > t_{1-\alpha/2,n-k}$ with the test statistics

$$T^{(j)} := \frac{\hat{\theta}_j}{\sqrt{\widehat{\sigma}^2[(X'X)^{-1}]_{jj}}}$$

- pvalue $= \mathbb{P}_{\mathcal{H}_0}\left(|T^{(j)}| > |T^{(j)}|^{obs}\right) = \mathbb{P}\left(|\mathcal{T}(n-k)| > |T^{(j)}|^{obs}\right)$

Example (multiple linear regression)

```r
summary(regmulti)
```

```

Call:
lm(formula = oxy ~ ., data = fitness)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.4026 -0.8991  0.0706  1.0496  5.3847 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.93448   12.40326   8.299 1.64e-08 ***
age          -0.22697    0.09984  -2.273  0.03224 *  
weight       -0.07418    0.05459  -1.359  0.18687    
runtime      -2.62865    0.38456  -6.835 4.54e-07 ***
rstpulse     -0.02153    0.06605  -0.326  0.74725    
runpulse     -0.36963    0.11985  -3.084  0.00508 ** 
maxpulse      0.30322    0.13650   2.221  0.03601 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.317 on 24 degrees of freedom
Multiple R-squared:  0.8487,	Adjusted R-squared:  0.8108 
F-statistic: 22.43 on 6 and 24 DF,  p-value: 9.715e-09
```

Prediction

- Based on the n previous observations, we may be interested with the prediction of the response of the model for a new point $X_0 = (1, X_0^{(1)}, \ldots, X_0^{(p)}) \in \mathcal{M}_{1k}(\mathbb{R})$:

$$Y_0 = X_0\theta + \varepsilon_0,\ \varepsilon_0 \sim \mathcal{N}(0, \sigma^2),\ \varepsilon_0 \amalg (\varepsilon_1, \ldots, \varepsilon_n)$$

- The predicted value is

$$\widehat{Y_0} = X_0\widehat{\theta} \sim \mathcal{N}(X_0\theta, \sigma^2 X_0(X'X)^{-1}X_0').$$

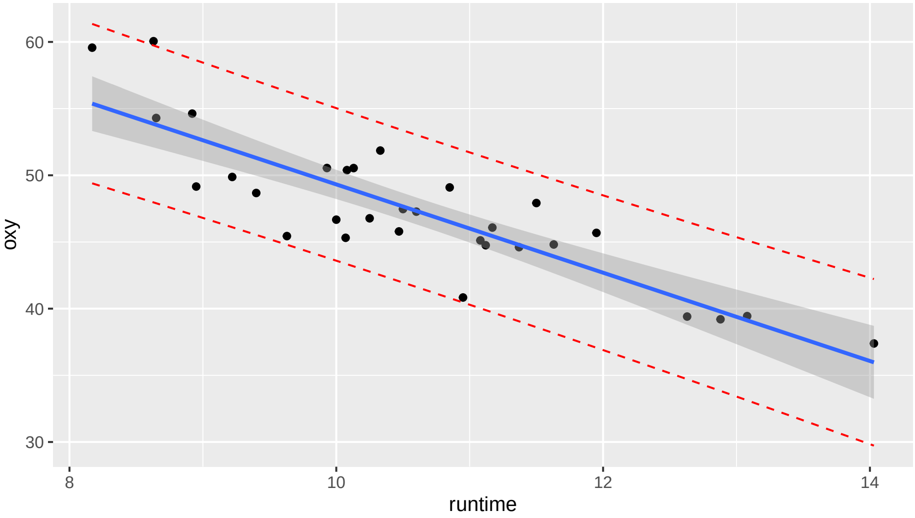
- Confidence interval for **the mean response** $X_0\theta$:

$$IC_{1-\alpha}(X_0\theta) = \left[\widehat{Y_0} \pm t_{1-\alpha/2,n-k} \times \widehat{\sigma}\sqrt{X_0(X'X)^{-1}X_0'}\ \right].$$

- Prediction interval for **the response** Y_0:

$$IC_{1-\alpha}(Y_0) = \left[\widehat{Y_0} \pm t_{1-\alpha/2,n-k} \times \widehat{\sigma}\sqrt{1 + X_0(X'X)^{-1}X_0'}\right].$$

Example



Measure for goodness-of-fit

- Decomposition of the variability:

$$SST = SSE + SSR$$

  - **Total sum of squares:** $SST = \|Y - \overline{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
  - **Explained sum of squares:** $SSE = \|\widehat{Y} - \overline{Y}\mathbb{1}_n\|^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$
  - **Residual sum of squares:**
    $SSR = \|Y - \widehat{Y}\|^2 = \sum_{i=1}^{n}(\widehat{\varepsilon}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$

- **Coefficient of determination:** proportion of explained variance

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{var(\widehat{Y})}{var(Y)} \in [0, 1]$$

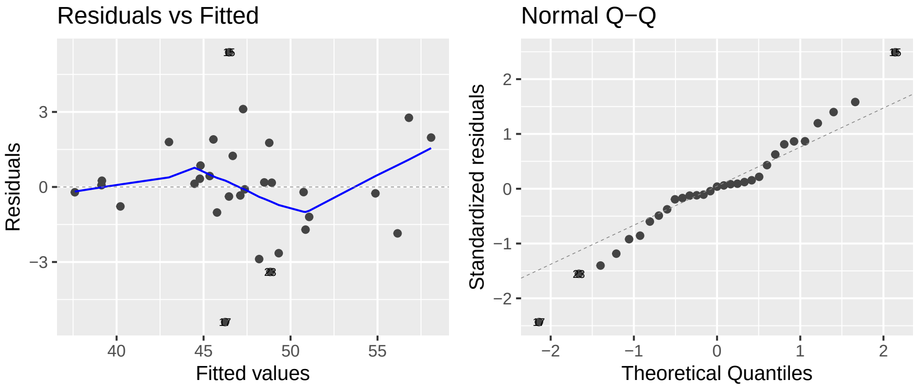
Example

```r
summary(regmulti)
```

```
Call:
lm(formula = oxy ~ ., data = fitness)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.4026 -0.8991  0.0706  1.0496  5.3847 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.93448   12.40326   8.299 1.64e-08 ***
age          -0.22697    0.09984  -2.273  0.03224 *  
weight       -0.07418    0.05459  -1.359  0.18687    
runtime      -2.62865    0.38456  -6.835 4.54e-07 ***
rstpulse     -0.02153    0.06605  -0.326  0.74725    
runpulse     -0.36963    0.11985  -3.084  0.00508 ** 
maxpulse      0.30322    0.13650   2.221  0.03601 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.317 on 24 degrees of freedom
Multiple R-squared:  0.8487,	Adjusted R-squared:  0.8108 
F-statistic: 22.43 on 6 and 24 DF,  p-value: 9.715e-09
```

Validation - Residuals

```r
autoplot(regmulti,label.size=2, which=1:2)
```



Fisher test of a sub-model

- Question: Is it possible to "simplify" the linear model?
- We consider two models:
  - (M_1): $Y = X\theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$
  - (M_0): $Y = Z\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$

  where (M_0) is a **sub-model** of M_1: $Im(Z) \subset Im(X)$

  (M_0 is a particular case of M_1, e.g some variables are removed)

  - $k_0 = dim(Im(Z)) < k = dim(Im(X))$
- We want to test M_0 against M_1 to explain the variable response Y

Fisher test of a sub-model

- Test statistics:

$$F = \frac{(SSR_0 - SSR_1)/(k - k_0)}{SSR_1/(n-k)} = \frac{\|X\widehat{\theta} - Z\widehat{\beta}\|^2/(k-k_0)}{\|Y - X\widehat{\theta}\|^2/(n-k)}$$

  with $SSR_0 = \|Y - Z\widehat{\beta}\|^2$ and $SSR_1 = \|Y - X\widehat{\theta}\|^2$.

- Under H_0, $F \sim \mathcal{F}(k - k_0, n - k)$ (Fisher's distribution law)

- Reject H_0 if $F > f_{1-\alpha}$ where $f_{1-\alpha}$ is the $(1-\alpha)$-quantile of $F(k - k_0, n - k)$.

Example

Question: Can we simplify the model by considering only the explanatory variables *age*, *runtime* and *runpulse*?

```
reg0<-lm(oxy~age+runtime+runpulse,data=fitness)
anova(reg0,regmulti)
```

```
Analysis of Variance Table

Model 1: oxy ~ age + runtime + runpulse
Model 2: oxy ~ age + weight + runtime + rstpulse + runpulse + maxpulse
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 160.83
2     24 128.84  3    31.993 1.9865 0.1429
```

Example - Null model

- By default in the `summary` of `lm()`, the Fisher test for the null model is given
- Null model = no explanatory variable is used to explain the response variable Y: $Y_i = \theta_0 + \varepsilon_i$
  $(\theta_1 = \cdots = \theta_p = 0)$
- For the null model, $\widehat{\theta_0} = \overline{Y}$ and $SSR_0 = SST$.
- Test statistics: $F = \frac{SST - SSR_1/(p+1-1)}{SSR_1/n-(p+1)} = \frac{SSE_1/p}{SSR_1/n-(p+1)}$

```
regnull<-lm(oxy~1,data=fitness)
anova(regnull,regmulti)
```

```
Analysis of Variance Table

Model 1: oxy ~ 1
Model 2: oxy ~ age + weight + runtime + rstpulse + runpulse + maxpulse
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     30 851.38
2     24 128.84  6    722.54 22.433 9.715e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variable selection

- When we have a large number of variables (p large), we cannot test all the sub-models
- Variable selection algorithms are used to obtain some sub-models to test
- In part 2, some variable selection algorithms and regularized regression procedures will be presented.

Section 3

ANOVA

Context - Notation

- ANOVA = analysis of variance
- Aim: Explain a **quantitative variable** Y using **qualitative** explanatory variables named **factors**
- The modalities of a factor = **levels** (sub-groups in the sample)
- Here we will not address the issue of **experimental design**

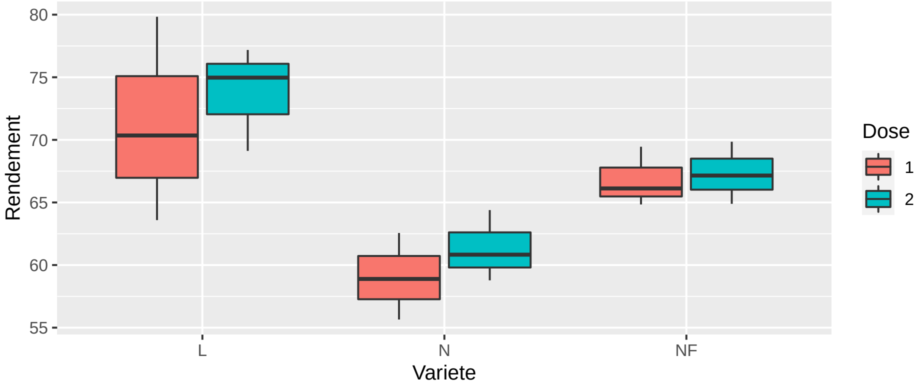
Example (wheat yield)

In a study of factors influencing wheat yield, three varieties of wheat (L, N and NF) and two nitrogen inputs were compared (normal supply = dose 1, intensive supply = dose 2).
Three repetitions for each couple (variety, dose) were performed and the yield (in q / ha) was measured.
We are interested in the differences that could exist from one variety to another, and in the possible interactions of varieties with nitrogen inputs.

```
 Dose   Variete    Rendement    
 1:9    L :6    Min.   :55.65  
 2:9    N :6    1st Qu.:62.82  
        NF:6    Median :65.50  
                Mean   :66.58  
                3rd Qu.:69.75  
                Max.   :79.83  
```

Example

```r
ggplot(Ble, aes(x=Variete, y=Rendement, fill=Dose)) +
  geom_boxplot()
```

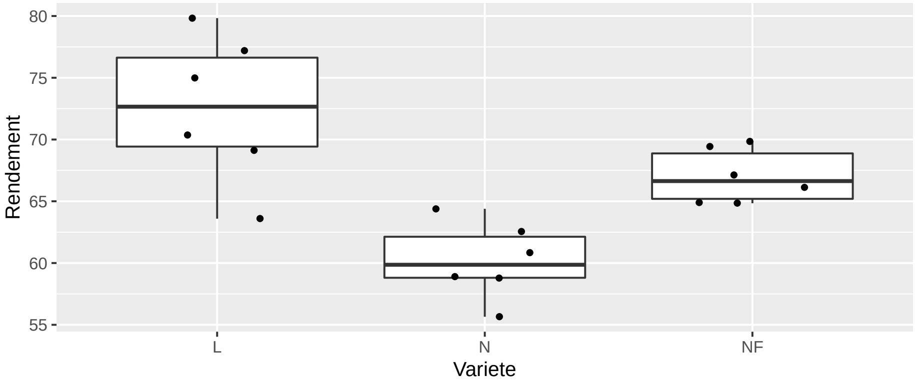


One-way ANOVA: Context

- Data: One quantitative response variable Y and **one** factor having I levels
- Notation:
  - Y_{ij} = value for individual j in group i (level of the factor)
  - Group i has n_i individuals
  - $Y_{i.}$ is the mean value for group i
  - $n = \sum_{i=1}^{I} n_i$ is the number of individuals all together
- Question: potential effect of the factor on the response Y
  $\Leftrightarrow$ Difference of the average response variable by group

Example

Variety (i)	L	N	NF
Yield Y_{ij}	70.35 63.59 79.83	62.56 58.89 55.65	69.45 64.84 66.12
	74.97 69.12 77.18	58.78 64.39 60.83	69.85 64.89 67.15
Number n_i	6	6	6
Average $Y_{i.}$	72.50667	60.18333	67.05



One-way ANOVA: regular model

- **Regular model**: $\forall i = 1, \cdots I,\ \forall j = 1, \cdots, n_i$,

$$Y_{ij} = m_i + \varepsilon_{ij},\ \varepsilon_{ij} \text{ i.i.d } \mathcal{N}(0, \sigma^2)$$

$$Y = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1_{n_1} & 0_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & 0_{n_2} & \cdots & 0_{n_2} \\ 0_{n_3} & 0_{n_3} & 1_{n_3} & \cdots & 0_{n_3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{n_I} & 0_{n_I} & 0_{n_I} & \cdots & 1_{n_I} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_I \end{pmatrix} + \varepsilon$$

- Unknown parameters: $\theta = (m_1, \ldots, m_I)'\ [k = I] + \sigma^2$

One-way ANOVA: regular model

- $X'X$ is invertible $\Rightarrow$ regular model
- $\hat{\theta} = (X'X)^{-1}X'Y$ thus $\widehat{m}_i = Y_{i.}$

```r
summary(lm(Rendement~Variete-1,data=Ble))
```

```
Call:
lm(formula = Rendement ~ Variete - 1, data = Ble)

Residuals:
   Min     1Q Median     3Q    Max 
-8.917 -2.159 -0.415  2.447  7.323 

Coefficients:
          Estimate Std. Error t value Pr(>|t|)    
VarieteL    72.507      1.664   43.58  < 2e-16 ***
VarieteN    60.183      1.664   36.17 5.21e-16 ***
VarieteNF   67.050      1.664   40.30  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.076 on 15 degrees of freedom
Multiple R-squared:  0.9969,    Adjusted R-squared:  0.9963 
F-statistic:  1610 on 3 and 15 DF,  p-value: < 2.2e-16
```

One-way ANOVA: singular model

- **Singular model:** $\forall i = 1, \cdots I, \ \forall j = 1, \cdots, n_i,$

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where

- μ = average effect
- $\alpha_i = m_i - \mu$ = differential effect of group i.
- But this model is over-parameterized [I+1 parameters]

  $\Rightarrow$ one constraint is required to have an identifiable model

  - Orthogonal constraint : $\sum_{i=1}^{I} n_i \alpha_i = 0$
  - By default in R: $\alpha_1 = 0$

One-way ANOVA: singular model

- Estimation of θ:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \theta = (\mu, \alpha_1, \dots, \alpha_I)'$$

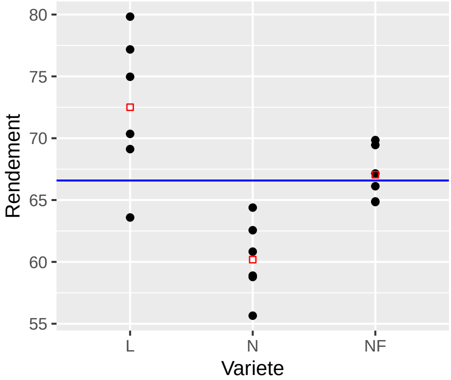
  - With the constraint $\sum_{i=1}^{I} n_i \alpha_i = 0$:

$$\begin{cases} \widehat{\mu} = Y_{..} \\ \widehat{\alpha}_i = Y_{i.} - Y_{..} \end{cases}$$

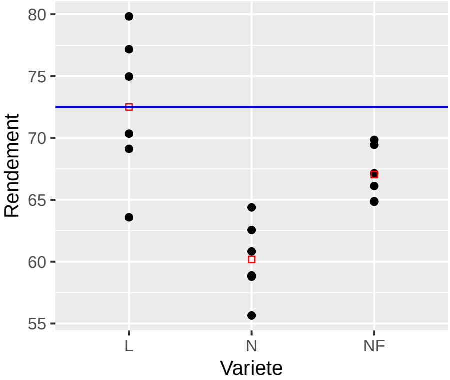
  - With the constraint $\alpha_1 = 0$:

$$\begin{cases} \widehat{\mu} = Y_{1.} \\ \widehat{\alpha}_i = Y_{i.} - Y_{1.} \end{cases}$$

Example



80
75
70
65
60
55
Rendement
L
N
NF
Variete



80
75
70
65
60
55
Rendement
L
N
NF
Variete

Example

```r
anov1 <- lm(Rendement~Variete,data=Ble)
summary(anov1)
```

```

Call:
lm(formula = Rendement ~ Variete, data = Ble)

Residuals:
   Min      1Q  Median      3Q     Max
-8.917 -2.159 -0.415  2.447  7.323

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.507      1.664  43.578   <2e-16 ***
VarieteN     -12.323      2.353  -5.237   0.0001 ***
VarieteNF     -5.457      2.353  -2.319   0.0349 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.076 on 15 degrees of freedom
Multiple R-squared:  0.6475,    Adjusted R-squared:  0.6005
F-statistic: 13.77 on 2 and 15 DF,  p-value: 0.0004018
```

One-way ANOVA: predictions, residuals and variance

- Predicted values:

$$\widehat{Y}_{ij} = \widehat{m}_i = \widehat{\mu} + \widehat{\alpha}_i = Y_{i.}$$

- Residuals:

$$\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{Y}_{ij} = Y_{ij} - Y_{i.}$$

- Estimator of the variance σ^2:

$$\begin{aligned} \widehat{\sigma^2} &= \frac{\|Y - \widehat{Y}\|^2}{n-I} = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{Y}_{ij})^2 \\ &= \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{\varepsilon}_{ij})^2 = \frac{SSR}{n-I} \end{aligned}$$

One-way ANOVA: effect of the factor?

- Testing procedure:

$$\mathcal{H}_0 : m_1 = m_2 = \cdots = m_I = m \Longleftrightarrow \forall i = 1, \cdots, I,\ \alpha_i = 0$$

versus

$$\mathcal{H}_1 : \exists (i, i') \text{ such that } m_i \neq m_{i'}.$$

- Fisher test of the sub-model:

$$(M_0) : Y_{ij} = m + \varepsilon_{ij} \text{ with } \widehat{m} = Y_{..} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}$$

```r
anmequal<-lm(Rendement~1,data=Ble)
anova(anmequal,anov1)
```

```
Analysis of Variance Table

Model 1: Rendement ~ 1
Model 2: Rendement ~ Variete
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     17 706.73
2     15 249.15  2    457.58 13.774 0.0004018 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-way ANOVA

- We are interested in the effect of **Variety** and **Dose** on **wheat yield** with a possible interaction between the two factors.
- Y_{ijk} = wheat yield of the k-th plot with Dose i and Variety i
- Two factors:
    - Factor A (*Dose*) with $I = 2$ levels
    - Factor B (*Variety*) with $J = 3$ levels
- n_{ij} = nb of obs. for level i of factor A and level j of factor B
- $Y_{ij.}$ = mean of observations in cell (i, j)
- Notation:

$$Y_{i..} = \frac{1}{n_{i+}} \sum_{j=1}^{J} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} \text{ with } n_{i+} = \sum_{j=1}^{J} n_{ij}$$

$$Y_{.j.} = \frac{1}{n_{+j}} \sum_{i=1}^{I} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} \text{ with } n_{+j} = \sum_{i=1}^{I} n_{ij}$$

$$Y_{...} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{\ell=1}^{n_{ij}} Y_{ij\ell} \text{ with } n = \sum_{i=1}^{I} n_{i+} = \sum_{j=1}^{J} n_{+j}$$

Two-way ANOVA: Example

Variety	L ($j = 1$)	N ($j = 2$)	NF ($j = 3$)
Dose 1 ($i = 1$)			
$(Y_{1,j,k})$	70.35 63.59 79.83	62.56 58.89 55.65	69.45 64.84 66.12
$n_{1j} = 3$	$Y_{11.} = 71.26$	$Y_{12.} = 59.03$	$Y_{13.} = 66.80$
Dose 2 ($i = 2$)			
$(Y_{2,j,k})$	74.97 69.12 77.18	58.78 64.39 60.83	69.85 64.89 67.15
$n_{2j} = 3$	$Y_{21.} = 73.76$	$Y_{22.} = 61.33$	$Y_{23.} = 67.30$

Two-way ANOVA: Models

- **Regular** model:

$$Y_{ijk} = m_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

  but all the effects are included in m_{ij}

- **Singular** model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

  distinction of effects ... but constraints are required for parameter estimation.
  ($1 + I + J + IJ$ parameters, IJ ddl thus $1 + I + J$ constraints)

Two-way ANOVA: Estimation and singular model

- Orthogonal constraints: If $n_{ij} = \frac{n_{i+}n_{+j}}{n}$, the orthogonal constraints (type I) are

$$\sum_{i=1}^{I} n_{i+}\alpha_i = 0; \sum_{j=1}^{J} n_{+j}\beta_j = 0; \forall i, \sum_{j=1}^{J} n_{ij}\gamma_{ij} = 0; \forall j, \sum_{i=1}^{I} n_{ij}\gamma_{ij} = 0.$$

$$\Longrightarrow \widehat{\mu} = Y_{...}, \widehat{\alpha}_i = Y_{i..} - Y_{...}, \widehat{\beta}_j = Y_{.j.} - Y_{...}, \widehat{\gamma}_{ij} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$$

- By default in R: $\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0$

$$\Longrightarrow \widehat{\mu} = Y_{11.}, \widehat{\alpha}_i = Y_{i1.} - Y_{11.}, \widehat{\beta}_j = Y_{1j.} - Y_{11.}, \widehat{\gamma}_{ij} = Y_{ij.} - Y_{i1.} - Y_{1j.} + Y_{11.}$$

Example

```
anov2 = lm(Rendement~Dose*Variete,data=Ble)
summary(anov2)
```

```
Call:
lm(formula = Rendement ~ Dose * Variete, data = Ble)

Residuals:
   Min     1Q Median     3Q    Max
-7.667 -2.296 -0.325  2.623  8.573

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       71.257      2.536  28.101 2.55e-12 ***
Dose2              2.500      3.586   0.697  0.49899
VarieteN         -12.223      3.586  -3.409  0.00519 **
VarieteNF         -4.453      3.586  -1.242  0.23801
Dose2:VarieteN    -0.200      5.071  -0.039  0.96919
Dose2:VarieteNF   -2.007      5.071  -0.396  0.69928
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.392 on 12 degrees of freedom
Multiple R-squared:  0.6725,    Adjusted R-squared:  0.536
F-statistic: 4.928 on 5 and 12 DF,  p-value: 0.01105
```

Predicted values, residuals and variance

- Predicted values:

$$\widehat{Y}_{ijk} = \widehat{m}_{ij} = Y_{ij.} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_{ij}$$

- Residuals:

$$\widehat{\varepsilon}_{ijk} = Y_{ijk} - Y_{ij.}$$

- Estimator of the variance σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n-IJ}\sum_{ijk}(\widehat{\varepsilon}_{ijk})^2 = \frac{1}{n-IJ}\sum_{ijk}(Y_{ijk} - Y_{ij.})^2 = \frac{SSR}{n-IJ}$$

Decomposition of the variability

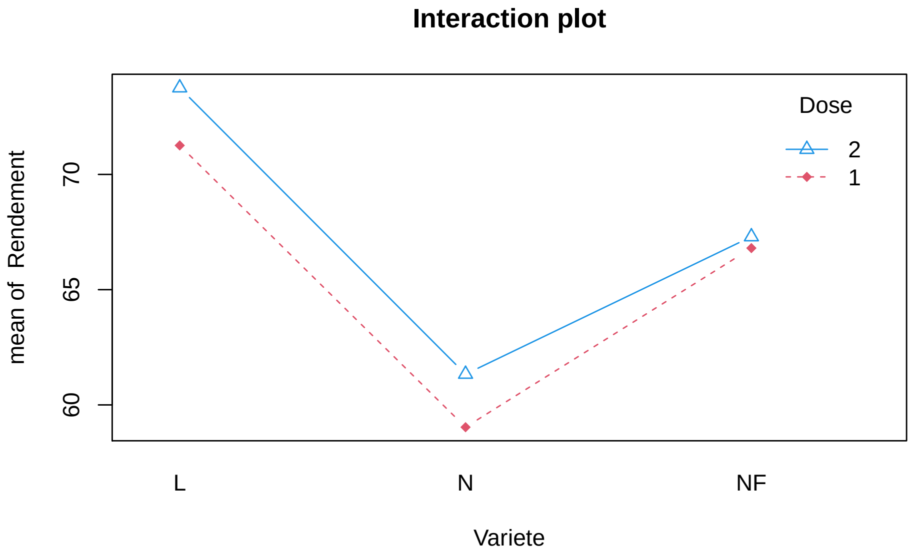
$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk}-Y_{...})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(Y_{ij.}-Y_{...})^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\, var_{ij}(Y)}_{\text{SSR}}$$

with $var_{ij}(Y) = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}(Y_{ijk}-Y_{ij.})^2$.

$$\begin{aligned}
SSE \quad &= \sum_{i=1}^{I} n_{i+}(Y_{i..}-Y_{...})^2 && = SSA \\
&+ \sum_{j=1}^{J} n_{+j}(Y_{.j.}-Y_{...})^2 && = SSB \\
&+ \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(Y_{ij.}-Y_{i..}-Y_{.j.}+Y_{...})^2 && = SSI
\end{aligned}$$

Two-way ANOVA: Interaction plot

```
attach(Ble)
interaction.plot(Variete,Dose,Rendement,col=c(2,4),pch=c(18,24),main="Interaction pl
```



Two-way ANOVA: non-interaction test

- Hypothesis:

$$\mathcal{H}_I : \gamma_{ij} = 0,\ \forall i = 1, \cdots, I,\ \forall j = 1, \cdots, J$$

- Fisher test of sub-model:
  - (M_1) (model with interaction): $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$
  - (M_0) (additive model): $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

- Test statistics:

$$F = \frac{SSI/(I-1)(J-1)}{SSR/(n-IJ)} \underset{\mathcal{H}_0}{\sim} \mathcal{F}((I-1)(J-1), n-IJ).$$

Example

```r
anov2add = lm(Rendement~Variete + Dose, data=Ble)
anova(anov2add,anov2)
```

```
Analysis of Variance Table

Model 1: Rendement ~ Variete + Dose
Model 2: Rendement ~ Dose * Variete
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     14 235.14
2     12 231.47  2    3.6654 0.095   0.91
```

Two-way ANOVA: Test for the factor effect

- Hypothesis: $\mathcal{H}_A : \alpha_i = 0, \ \forall i = 1, \cdots, I$
- Fisher testing of sub-model:
  - (M_1) (additive model): $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$
  - (M_0) (one-way ANOVA): $Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$
- Test statistics:

$$F = \frac{SSA/(I-1)}{SSRAB/(n-(I+J-1))} \underset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-(I+J-1)),$$

  where $SRAB =$ residual sum of squares for the additive model.

- Same for testing the absence of effect of factor B

Example

```r
anovA = lm(Rendement~Variete,data=Ble)
anova(anovA,anov2add)
```

```
Analysis of Variance Table

Model 1: Rendement ~ Variete
Model 2: Rendement ~ Variete + Dose
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     15 249.15
2     14 235.14  1     14.01 0.8341 0.3765
```

```r
anovB = lm(Rendement~Dose,data=Ble)
anova(anovB,anov2add)
```

```
Analysis of Variance Table

Model 1: Rendement ~ Dose
Model 2: Rendement ~ Variete + Dose
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     16 692.72
2     14 235.14  2    457.58 13.622 0.0005192 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 4

ANCOVA

Context

- ANCOVA= Analysis of covariance
- We want to explain a **quantitative** response variable Y using **qualitative** and **quantitative** variables together
- Here we only consider one covariate z and one factor T with I levels

Example

We want to find if temperature and oxygenation conditions influence the evolution of oyster weight. We have $n = 20$ bags of 10 oysters. We place, during a month, these 20 bags randomly in $I = 5$ different locations of a channel cooling of a power station at the rate of $n_i = 4$ bags per location. These locations are differentiated by their temperatures and oxygenations.

For each bag, we have

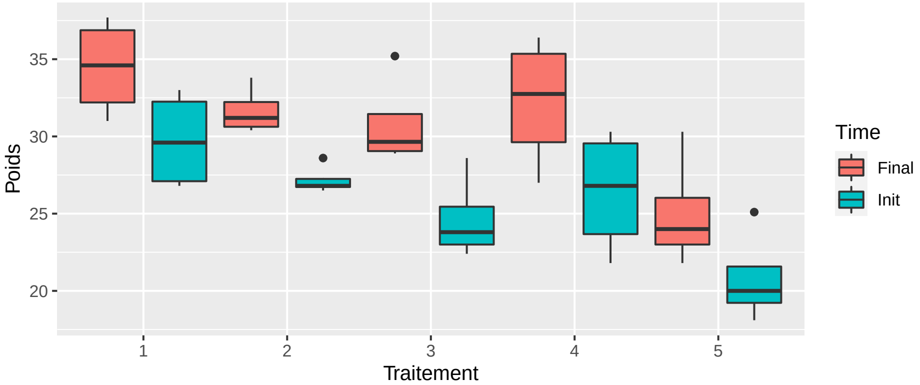
- its weight after the experiment (*Pds Final*) = the response Y
- its weight before the experiment (*Pds Init*) = the explanatory variable z
- the location (*Treatment* - 1 to 5) = the qualitative variable T

Example

```r
print(Huitres)
```

```
   PdsInit PdsFinal Traitement
1     27.2     32.6          1
2     32.0     36.6          1
3     33.0     37.7          1
4     26.8     31.0          1
5     28.6     33.8          2
6     26.8     31.7          2
7     26.5     30.7          2
8     26.8     30.4          2
9     28.6     35.2          3
10    22.4     29.1          3
11    23.2     28.9          3
12    24.4     30.2          3
13    29.3     35.0          4
14    21.8     27.0          4
15    30.3     36.4          4
16    24.3     30.5          4
17    20.4     24.6          5
18    19.6     23.4          5
19    25.1     30.3          5
20    18.1     21.8          5
```

Example



Regular model

- Model:

$$(MR) : \begin{cases} Y_{ij} = a_i + b_i \ z_{ij} + \varepsilon_{ij}, & \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i \\ \varepsilon_{ij} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

$\Leftrightarrow$ Estimating a linear regression of Y on z for each level i of the factor T.

$$\begin{pmatrix} Y_{(1)} \\ \vdots \\ \vdots \\ Y_{(I)} \end{pmatrix} = \begin{pmatrix} X_{(1)} & & & \\ & X_{(2)} & & \\ & & \ddots & \\ & & & X_{(I)} \end{pmatrix} \begin{pmatrix} a_1 \\ b_1 \\ \vdots \\ a_I \\ b_I \end{pmatrix} + \begin{pmatrix} \varepsilon_{(1)} \\ \vdots \\ \vdots \\ \varepsilon_{(I)} \end{pmatrix}$$

with $Y_{(i)} = (Y_{i1}, \ldots, Y_{in_i})'$, $X_{(i)} = \left(\mathbb{1}_{n_i}, z_{(i)}\right)$.

Singular model

$$(MS) : \begin{cases} Y_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i) z_{ij} + \varepsilon_{ij}, & \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i. \\ \varepsilon_{ij} \text{ i.i.d } \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

- In this parametrization,
  - interaction effect between the covariate z and the factor T: γ_i
  - differential effect of the factor T on Y: α_i
  - differential effect of the covariate z on Y: β
- $2I + 2$ parameters $\Rightarrow$ 2 constraints are required to model identifiability

Parameter estimation

- **Regular model:**

$$\hat{\theta} = \begin{pmatrix} \widehat{a_1} \\ \widehat{b_1} \\ \vdots \\ \widehat{a}_I \\ \widehat{b}_I \end{pmatrix} = \begin{pmatrix} (X'_{(1)} X_{(1)})^{-1} X'_{(1)} Y_{(1)} \\ \vdots \\ (X'_{(I)} X_{(I)})^{-1} X'_{(I)} Y_{(I)} \end{pmatrix}$$

- **Singular model** with constraints $\alpha_1 = \gamma_1 = 0$ (default in R):

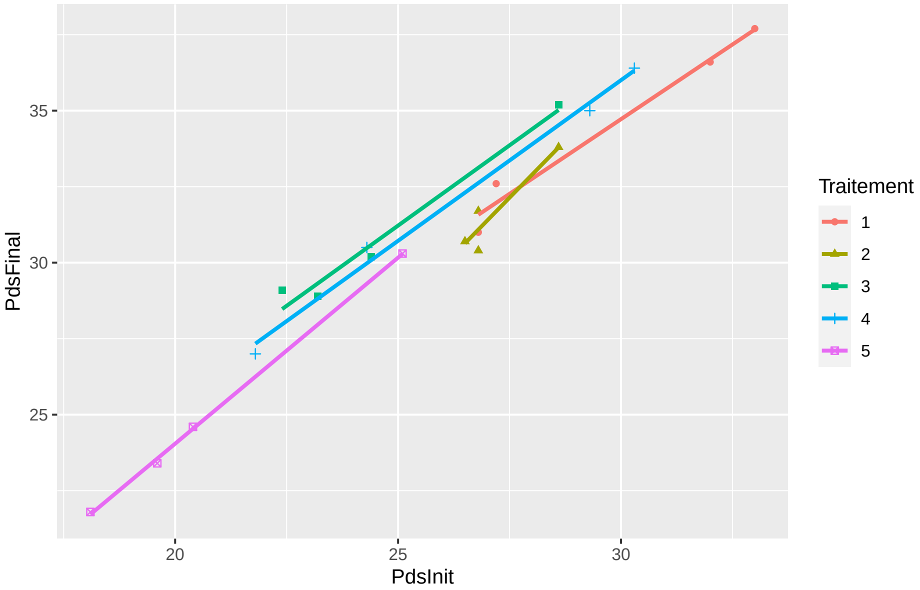
$$\begin{cases} \widehat{\mu} = \widehat{a}_1 \\ \widehat{\alpha}_i = \widehat{a}_i - \widehat{a}_1 \\ \widehat{\beta} = \widehat{b}_1 \\ \widehat{\gamma}_i = \widehat{b}_i - \widehat{b}_1 \end{cases}$$

Example

```
complet<-lm(PdsFinal~PdsInit * Traitement)
summary(complet)
```

```
Call:
lm(formula = PdsFinal ~ PdsInit * Traitement)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68699 -0.28193  0.02184  0.10425  0.63075

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.24126    2.86473   1.830   0.0972 .
PdsInit              0.98265    0.09588  10.249 1.27e-06 ***
Traitement2        -14.39058    9.15971  -1.571   0.1472
Traitement3         -0.42330    3.97747  -0.106   0.9174
Traitement4         -0.94550    3.50725  -0.270   0.7930
Traitement5         -5.67309    3.57150  -1.588   0.1433
PdsInit:Traitement2  0.51871    0.33406   1.553   0.1515
PdsInit:Traitement3  0.07342    0.14699   0.499   0.6282
PdsInit:Traitement4  0.07428    0.12229   0.607   0.5571
PdsInit:Traitement5  0.24124    0.13980   1.726   0.1151
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5324 on 10 degrees of freedom
Multiple R-squared:  0.9921,	Adjusted R-squared:  0.985
F-statistic: 139.5 on 9 and 10 DF,  p-value: 2.572e-09
```

Example



Test of non-interaction between factor and covariate

We want to test the null hypothesis:

$$\mathcal{H}_0^{(SI)} : b_1 = b_2 = \cdots = b_I \Longleftrightarrow \gamma_1 = \gamma_2 = \cdots = \gamma_I = 0$$

Fisher's test to compare the complete model

$$(MS): \ \ Y_{ij} = (\mu + \alpha_i) + (\beta + \gamma_i)z_{ij} + \varepsilon_{ij}, \, \forall i = 1, \cdots, I, \, \forall j = 1, \cdots, n_i$$

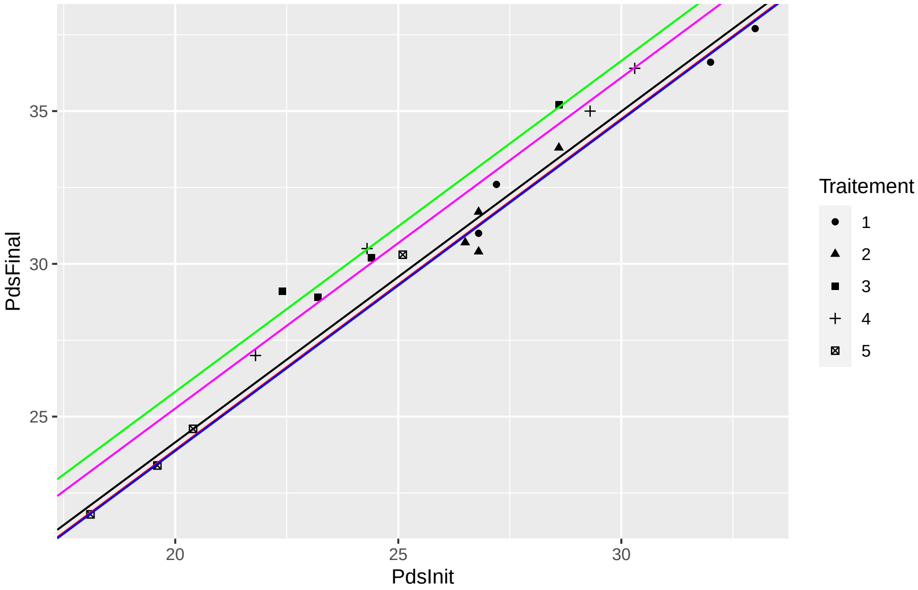
and the sub-model with non-interaction:
$$(MSI): \ \ Y_{ij} = \mu + \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \, \forall i = 1, \cdots, I, \, \forall j = 1, \cdots, n_i.$$

Test of non-interaction between factor and covariate

```r
nonI<-lm(PdsFinal~PdsInit+Traitement)
anova(nonI,complet)
```

```
Analysis of Variance Table

Model 1: PdsFinal ~ PdsInit + Traitement
Model 2: PdsFinal ~ PdsInit * Traitement
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     14 4.2223
2     10 2.8340  4    1.3883 1.2247 0.3602
```

Test of non-interaction between factor and covariate



ANCOVA with no-interaction

If the model with non-interaction between the factor and the covariate is retained

- Singular model:

$$Y_{ij} = \mu + \alpha_i + \beta\ z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

- Regular model:

$$Y_{ij} = a_i + b\ z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

we may test the effect of the factor or the effect of the covariate on the response.

Effect of the covariate z on Y

Fisher's test to compare the model with non-interaction

$$(M1): \; Y_{ij} = \mu + \alpha_i + \beta z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

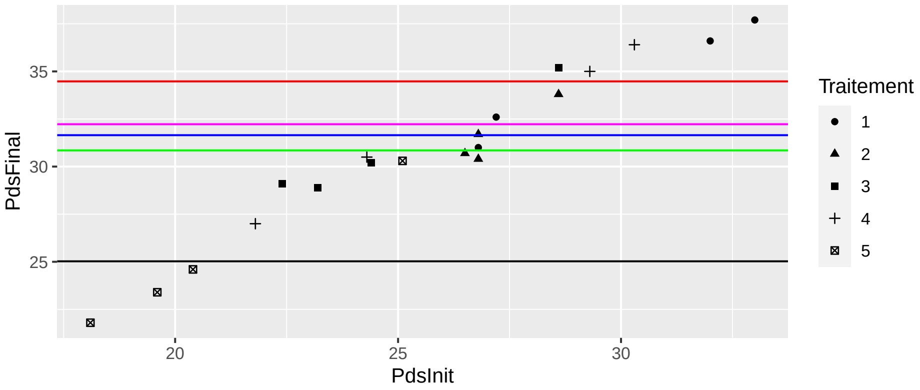
and the one-way ANOVA

$$(M2): \; Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i.$$

```
M2<-lm(PdsFinal~Traitement)
anova(M2,nonI)
```

```
Analysis of Variance Table

Model 1: PdsFinal ~ Traitement
Model 2: PdsFinal ~ PdsInit + Traitement
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     15 160.263
2     14   4.222  1    156.04 517.38 1.867e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effect of the covariate z on Y



Effect of the factor T on Y

Fisher's test to compare the model with non-interaction

$$(M1): \;\; Y_{ij} = \mu + \alpha_i + \beta\ z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

and the linear regression

$$(M3): \;\; Y_{ij} = \mu + \beta\ z_{ij} + \varepsilon_{ij}, \forall i = 1, \cdots, I, \forall j = 1, \cdots, n_i$$

```r
M3<-lm(PdsFinal~PdsInit)
anova(M3,nonI)
```

```
Analysis of Variance Table

Model 1: PdsFinal ~ PdsInit
Model 2: PdsFinal ~ PdsInit + Traitement
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 16.3117
2     14  4.2223  4    12.089 10.021 0.0004819 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effect of the factor T on Y