MUSÉUM NATIONAL D'HISTOIRE NATURELLE

SEMIN-R

Le modèle linéaire avec R : fonction lm()

Sébastien BALLESTEROS

UMR 7625 Ecologie Evolution
Ecole Normale Supérieure
46 rue d'Ulm
F-75230 Paris Cedex 05

balleste@biologie.ens.fr

1) un premier aperçu autour de la régression linéaire

Présentation

On cherche à décrire la relation entre le Taux de DDT d'un brochet (variable à expliquer y) et l'âge du brochet (variable explicative x)

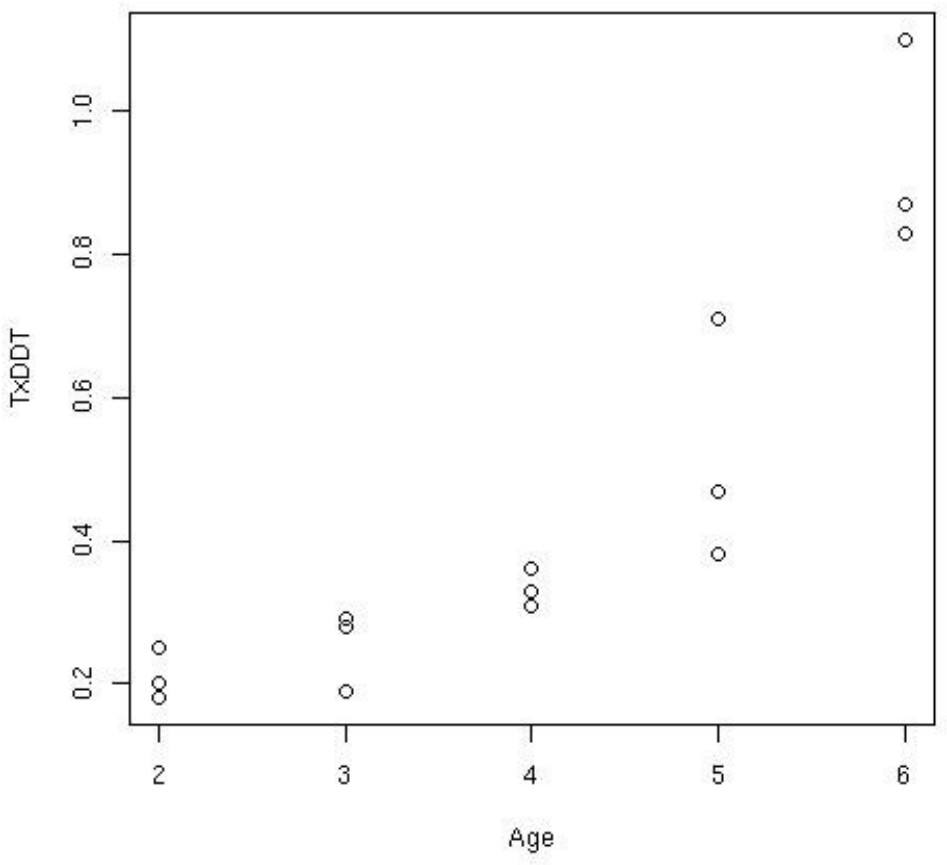
Données

On dispose d'un echantillon de n=15 brochets. Pour chaque brochet, on a
•son âge
•la mesure de son tx de DDT

```
> brochets<-read.table("ddt.dat", header = T)
> brochets
```

```
   Obs Age TxDDT
1    1   2  0.20
2    2   2  0.25
3    3   2  0.18
4    4   3  0.19
5    5   3  0.29
6    6   3  0.28
7    7   4  0.31
8    8   4  0.33
9    9   4  0.36
10  10   5  0.71
11  11   5  0.38
12  12   5  0.47
13  13   6  1.10
14  14   6  0.87
15  15   6  0.83
```

```
> attach(brochets)
> plot(Age,TxDDT)
```



écriture du modèle

Modèle

$Y_i = a + bx_i + E_i \qquad \{E_i\}\ i.i.d. \sim N(0,\sigma^2)$, i=1,...,n=15

- Indice i : n° brochet
- Variable Y_i Taux de DDT du i-ème brochet
- Variable x_i âge du i-ème brochet
- Variable E_i terme résiduel aléatoire
- σ^2 variance résiduelle
- a et b paramètres inconnus

Ecriture en terme de loi des Y_i

$Y_i \sim N(\mu_i, \sigma^2), \{Y_i\}$ independants

en notant $\mu_i = a + bx_i$

Sous R : lm(variable à expliquer ~ variable(s) explicative(s), ...)

>model<-lm(TxDDT~Age, data = brochets)

⟷

>model<-lm(brochets$TxDDT~brochets$Age)

⟷

>model<-lm(TxDDT~Age) si TxDDT et Age sont défini (avec par ex attach(brochets)

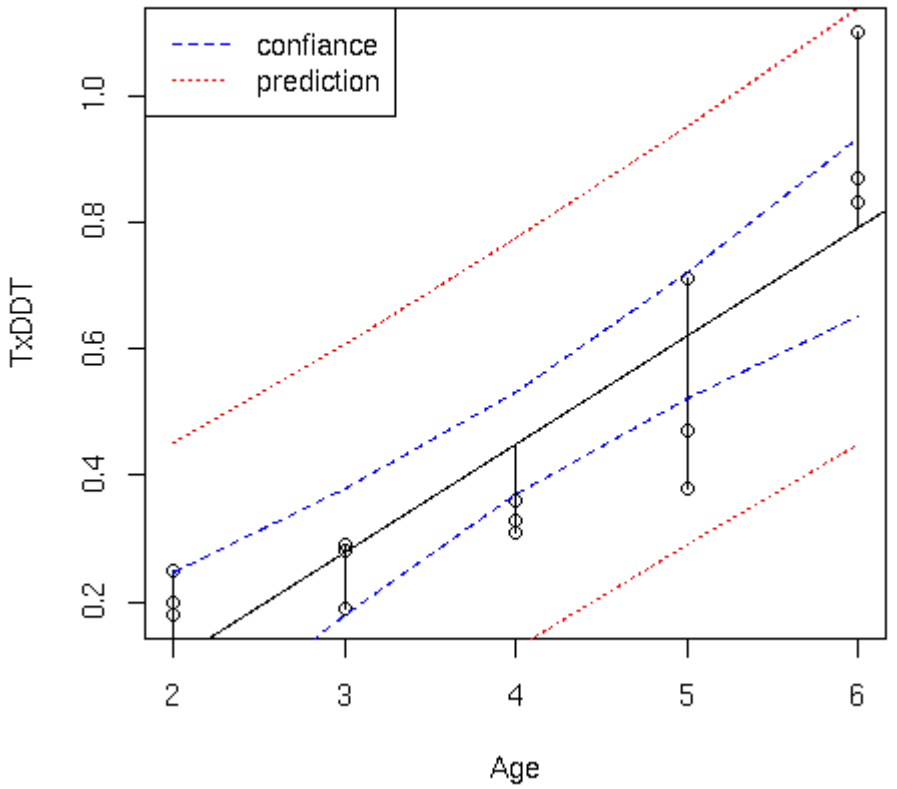
données manquantes : na.action

>model<-lm(TxDDT~Age, na.action =na.omit) par defaut

>model<-lm(TxDDT~Age, na.action =na.fail) produit un avertissement si NA(s)

l'objet lm sous R

```
> model<-lm(TxDDT~Age)
> model

Call:
lm(formula = TxDDT ~ Age)

Coefficients:
(Intercept)          Age
    -0.2353       0.1713

> coef(model)

(Intercept)         Age
 -0.2353333   0.1713333

> fitted(model)

        1         2         3         4         5         6         7         8
0.1073333 0.1073333 0.1073333 0.2786667 0.2786667 0.2786667 0.4500000
0.4500000
        9        10        11        12        13        14        15
0.4500000 0.6213333 0.6213333 0.6213333 0.7926667 0.7926667 0.7926667

> residuals(model)

          1           2           3           4           5           6
 0.092666667  0.142666667  0.072666667 -0.088666667  0.011333333
0.001333333
          7           8           9          10          11          12
-0.140000000 -0.120000000 -0.090000000  0.088666667 -0.241333333
-0.151333333
         13          14          15
 0.307333333  0.077333333  0.037333333

> predict(model, interval="confidence")

        fit        lwr       upr
1  0.1073333 -0.03322394 0.2478906
2  0.1073333 -0.03322394 0.2478906
3  0.1073333 -0.03322394 0.2478906
4  0.2786667  0.17927767 0.3780557
5  0.2786667  0.17927767 0.3780557
6  0.2786667  0.17927767 0.3780557
[...]
```




```
plot(Age,TxDDT)
abline(model)
segments(Age,fitted(model),Age, TxDDT)

pred.frame<-data.frame(Age=2:6)
pc<-predict(model, interval="confidence",
newdata=pred.frame)
pp<-predict(model, interval="prediction",
newdata=pred.frame)
matlines(pred.frame, pc[,2:3], lty=c(2,2), col="blue")
matlines(pred.frame, pp[,2:3], lty=c(3,3), col="red")
legend("topleft",c("confiance","prediction"),lty=c(2,3)
, col=c("blue","red"))
```

summary()

$$Y_i = a + bx_i + E_i$$

```
> model<-lm(TxDDT~Age)
> summary(model)
```

```
Call:
lm(formula = TxDDT ~ Age)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.24133 -0.10500  0.01133  0.08300  0.30733
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23533    0.11269  -2.088    0.057 .
Age          0.17133    0.02656   6.450 2.16e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1455 on 13 degrees of freedom
Multiple R-Squared: 0.7619,     Adjusted R-squared: 0.7436
F-statistic: 41.61 on 1 and 13 DF,  p-value: 2.165e-05
```

rappel : a et b minimise la somme des résidus²

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\hat{Y}_i = \hat{a} + \hat{b}\, x_i$$

Estimation des paramètres et test sur les paramètres

$$Y_i = a + bx_i + E_i$$

```
> model<-lm(TxDDT~Age)
> summary(model)

Call:
lm(formula = TxDDT ~ Age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24133 -0.10500  0.01133  0.08300  0.30733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23533    0.11269  -2.088    0.057 .
Age          0.17133    0.02656   6.450 2.16e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[...]

Pour chaque paramètre, la table donne la statistique observée ainsi que la proba critique associées au test d'hypothese :

H0{le paramètre est nul} vs H1{le paramètre n'est pas nul}

paramètres	Paramètres estimés	Ecart type	Statistique de test	Probabilité critique
constante	$\hat{a}$	$\hat{\sigma}_a$	$T_a = \hat{a}/\hat{\sigma}_a$	Pc_a
pente	$\hat{b}$	$\hat{\sigma}_b$	$T_b = \hat{b}/\hat{\sigma}_b$	Pc_b

Ajustement du modèle

> summary(model)

Call:
lm(formula = TxDDT ~ Age)

[...]

Residual standard error: 0.1455 on 13 degrees of freedom
Multiple R-Squared: 0.7619, Adjusted R-squared: 0.7436
F-statistic: 41.61 on 1 and 13 DF, p-value: 2.165e-05

Estimation de σ^2

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$\hat{\sigma}^2 = SSR/n-2$$

R^2 : coefficient de determination, avant de le définir il faut définir 3 sommes de carrés.

On partitionne la variation totale de Y (SST) en 2 composantes évoquant le pouvoir explicatif de notre modèle : la variation expliquée par notre modèle (SSM) et la variation inexpliquée (SSR) -> SST = SSM+SSR

R^2 = SSM/SST = 1-SSR/SST
$0 \leq R^2 \leq 1$
Plus cette valeure sera proche de 1 meilleur sera l'ajustement.
S'interprete comme la proportion de variabilité de Y expliquée par le modèle.

R^2 ajusté : ajustement du R^2 au nombre p de variables explicatives.
R^2 ajusté = 1-(SSR/(n-p))/(SST(n-1))

Table d'analyse de la variance et test de modèles emboîtés

```
> model<-lm(TxDDT~Age)
> anova(model)   ⟺   summary.aov(model)
Analysis of Variance Table

Response: TxDDT
          Df  Sum Sq Mean Sq F value    Pr(>F)
Age        1 0.88065 0.88065  41.609 2.165e-05 ***
Residuals 13 0.27515 0.02117
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source	Degrés de liberté	Somme de carrés	Carré moyen	Statistique de test	Probabilité critique
Modèle	p-1	$SSM=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$	SSM/(p-1)	$F=\frac{SSM/(p-1)}{SSR/(n-p)}$	$P(F_{p-1,n-p}>F)$
Résidu	n-p	$SSR=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	SSR/(n-p)		
Total	n-1	$SST=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$	SST/(n-1)		

Test du *modèle constant* contre le *modèle complet*

$H0\ \{Y_i=a+E_i\}$ vs $H1\{Y_i=a+bx_i+E_i\}$

remarque : test équivalent au test de H0 = {b=0} en effet $F=((B/S_b)^2)$

bilan

$$Y_i = a + bx_i + E_i$$
$$\{E_i\}\, i.i.d \sim N(0, \sigma^2)$$

```
> model<-lm(TxDDT~Age)

> summary.aov(model)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Age          1 0.88065 0.88065  41.609 2.165e-05 ***
Residuals   13 0.27515 0.02117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(model)

Call:
lm(formula = TxDDT ~ Age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24133 -0.10500  0.01133  0.08300  0.30733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23533    0.11269  -2.088    0.057 .
Age          0.17133    0.02656   6.450 2.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1455 on 13 degrees of freedom
Multiple R-Squared: 0.7619,     Adjusted R-squared: 0.7436
F-statistic: 41.61 on 1 and 13 DF,  p-value: 2.165e-05
```

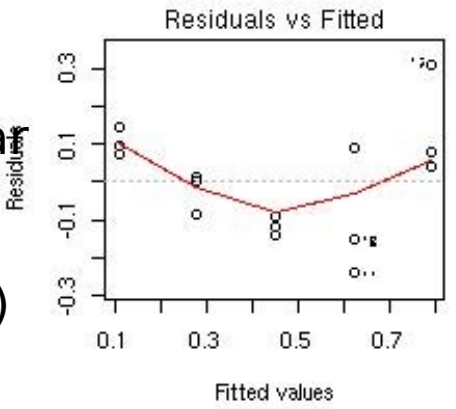
Validation du modèle

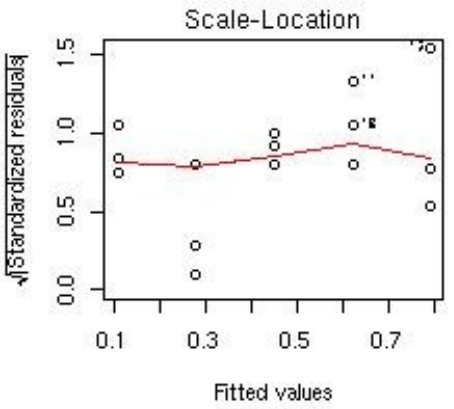
$Y_i \sim N(\mu_i, \sigma^2), \{Y_i\}$ independants

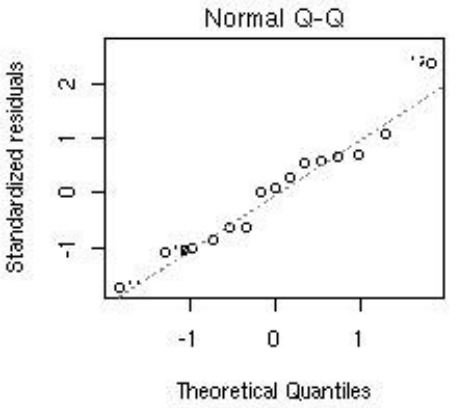
en notant $\mu_i = a + bx_i$

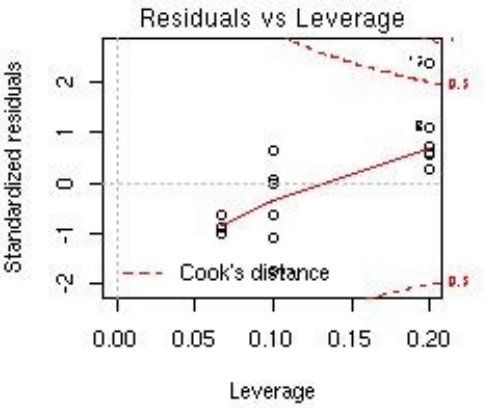
Le modèle suppose que la régression est linéaire, les termes d'erreurs ont même variance, ils sont indépendants et issus d'une loi normale.
Avant d'interpréter les résultats il est **indispensable** de vérifier ces hypothèses.

- linéarité de la relation
- Homoscédasticité / hétéroscédasticité
- Indépendance
- Normalité (hypothèse la moins importante car le modele linéaire est robuste à la normalité et les résidus suivent asymptotiquement une loi normale (i.e pour des grands echantillons))
- Points aberrants










```
> layout(matrix(1:4,2,2))
> plot(model)

>model2<-lm(log(TxDDT)~Age)
```

extraire des informations depuis l'objet model

rappel :

```
>coef(model)
>fitted(model)
>resid(model)
```

exemple avec test d'hypothèse pour voir si la pente est égale à 0.2

```
> summary(model)[[4]]
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.2353333 0.11269018 -2.088322 5.700736e-02
Age          0.1713333 0.02656133  6.450480 2.164828e-05
```

t=|0.1713333-0.2|/0.02656133

```
> t<-abs(summary(model)[[4]][[2]]-0.2)/summary(model)[[4]][[4]]
> 1-pt(t,13)
[1] 0.1500441
```

exemple de calcul de puissance : detecter la linearité.

```
nsim=400
pval<-numeric(nsim)

for (i in 1:nsim) {
    y_det = a + b * x
    y = rnorm(N, mean = y_det, sd = sd)
    m = lm(y ~ x)
    pval[i] = summary(m)[[4]][[8]]
}

sum(pval<0.05)/nsim
```

2)le modèle linéaire avec R

Le modèle linéaire avec R

régression linéaire, *ANOVA* sont des cas particuliers d'un **même** modèle statistique, le **modèle linéaire** que l'on peut écrire :

$$y_{tj} = m_t + e_{tj}$$

partie fixe, linéaire

partie aléatoire, normale

- quantitative → régression linéaire
- qualitative → ANOVA
- quantitative et qualitative → ANCOVA

erreurs indépendantes entres elles, suivant chacune une loi normale d'espérance nulle et de même variance.

- t est l'indice d'un traitement. Les differents facteurs pouvant intervenir dans sa définition sont contôlés, ils sont donc fixes, non aléatoires.
- j est un indice de répétition (pouvant ne pas exister explicitement)
- m_t est l'espérance de y_{tj}

$$y_{tj} \sim N(m_t, \sigma^2),\ \{y_{tj}\} \text{ indépendants}$$

Sous R : lm(variable à expliquer ~ variable(s) explicative(s), ...)

Le modèle linéaire avec R

definition du modèle
> model<-lm(variable à expliquer ~ variables explicative(s), ...)

table d'analyse de la variance, test de modèles emboités
> summary.aov(model)
ou >anova(model)

Estimation des parametres et tests sur les parametres / R^2 et S^2
> summary.lm(model)
ou >summary(model)

diagnostique (validation des hypothèses du modèle)
>plot(model)

3)ANOVA à un facteur

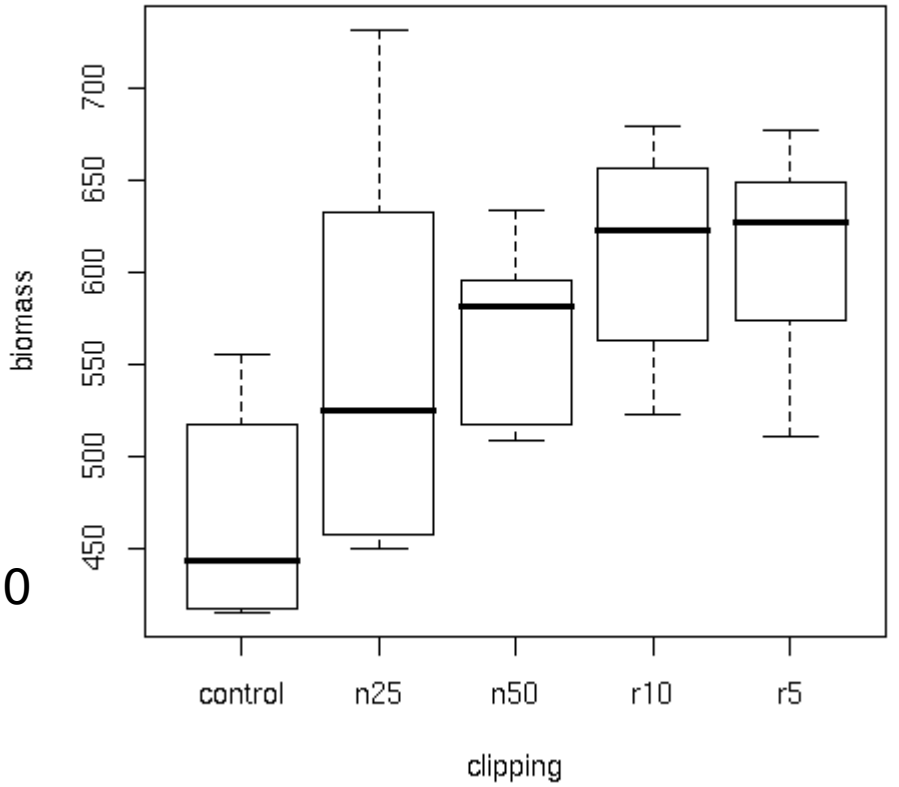
présentation

experience de competition chez des plantes

données

```
> results<-read.table("compet.txt",header=T
> results
   biomass clipping
1      551      n25
2      457      n25
3      450      n25
4      731      n25
5      499      n25
6      632      n25
7      595      n50
8      580      n50
9      508      n50
10     583      n50
11     633      n50
12     517      n50
13     639       r5
14     615       r5
15     511       r5
16     573       r5
17     648       r5
18     677       r5
19     417  control
20     449  control
21     517  control
22     438  control
23     415  control
24     555  control
25     563      r10
26     631      r10
27     522      r10
28     613      r10
29     656      r10
30     679      r10
```

control
shoot clipping treatment n25 & n50
root clipping treatment r5 & r10




```
>plot(biomass~clipping)
```

écriture du modèle

Modèle

$$Y_{ij} = \mu + \alpha_i + E_{ij} \qquad \{E_{ij}\} \; i.i.d. \sim N(0, \sigma^2)$$

- Indice i : represente le type de traitements (control, n25 & n50, r5 & r10)
- Indice j: est le numero de la parcelle au sein de son type
- Variable Y_{ij} biomasse de la j-ème parcelle du i-ème type
- Le paramètre μ est un terme constant
- Le parametre α_i est l'effet (additif) du traitement i
- Variable E_i terme résiduel aléatoire
- σ^2 variance résiduelle

Ecriture en terme de loi des Y_{ij}

$$Y_{ij} \sim N(\mu_i, \sigma^2), \{Y_{ij}\} \text{ independants}$$

en notant $\mu_i = \mu + \alpha_i$

Sous R : lm(variable à expliquer ~ variable(s) explicative(s), ...)

```
model<-lm(biomass~clipping)
model<-aov(biomass~clipping)
```

identique, differe par le summary affiché par défaut

Validation du modèle

$Y_{ij} \sim N(\mu_i, s^2), \{Y_{ij}\}$ independants

```
> layout(matrix(1:4,2,2))
> plot(model)
```

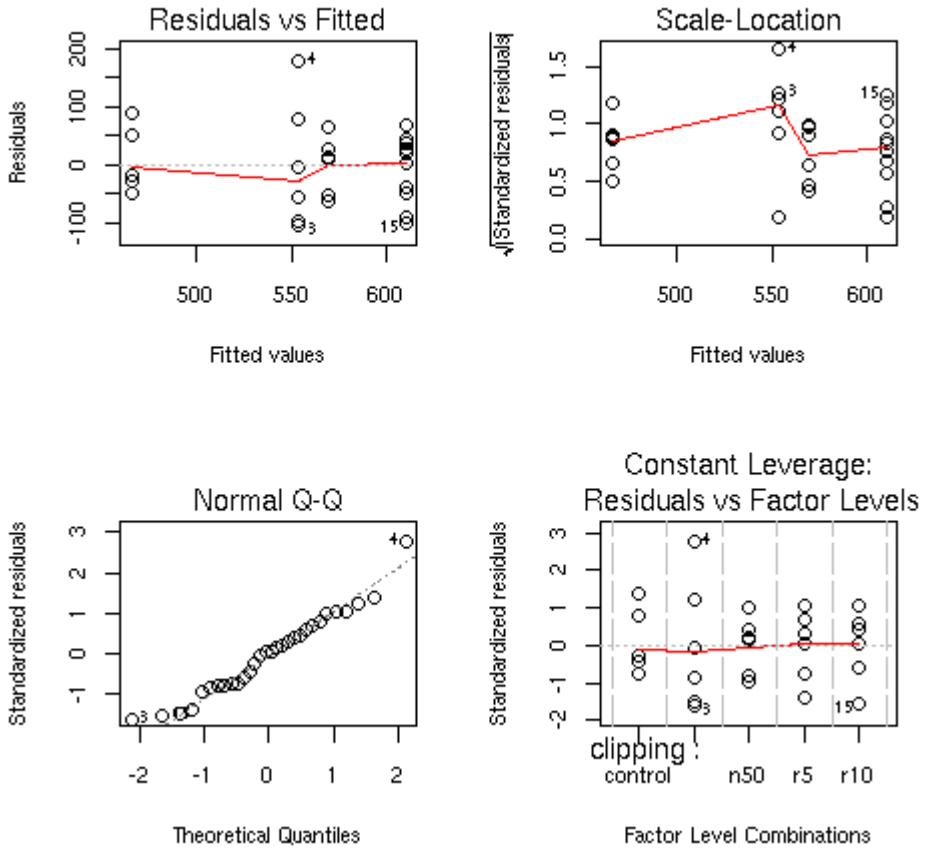


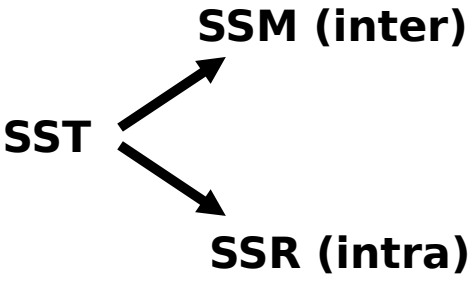
Table d'analyse de la variance et test de modèles emboîtés

Le facteur a t'il un effet (une influence) sur la variabilité de la variable étudiée ?
On étudie ici l'**effet du facteur (ici clipping) en générale** et non par exemple l'effet de 'n25'

```
> model<-lm(biomass~clipping)
> summary.aov(model)
```

ou bien

```
>model<-aov(biomass~clipping)
>summary(model)
```



$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{\bar{Y}})^2$$

$$SSR = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2$$

$$SSM = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{Y}_i - \bar{\bar{Y}})^2$$

```
            Df Sum Sq Mean Sq F value   Pr(>F)
clipping     4  85356   21339  4.3015 0.008752 **
Residuals   25 124020    4961
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source	Degrés de liberté	Somme de carrés	Carré moyen	Statistique de test	Probabilité critique
Modèle	k-1	SSM	SSM/(k-1)	$F = \frac{SSM/(k-1)}{SSR/(n-k)}$	$P(F_{k-1,n-k} > F)$
Résidu	n-k	SSR	SSR/(n-k) $= \hat{S}^2$		
Total	n-1	SST	SST/(n-1)	$F = \frac{\text{variance expliquee par le sol}}{\text{variance residuelle}}$	

Test du *modèle constant* contre le *modèle complet*

$H0 = \{Y_{ik} = \mu + E_{ij}\} = \{\alpha_1 = \alpha_2 = \alpha_3 = 0\}$ vs $H1 = \{Y_{ik} = \mu + \alpha_i + E_{ij}\} = \{\exists\, i : \alpha_i \neq 0\}$

Estimation des paramètres et test sur les paramètres

$$Y_{ij} = \mu + \alpha_i + E_{ij} \qquad \{E_{ij}\}\ i.i.d. \sim N(0, \sigma^2)$$

```
> model<-lm(biomass~clipping)
> summary(model)
[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   465.17      28.75  16.177  9.4e-15 ***
clippingn25    88.17      40.66   2.168  0.03987 *
clippingn50   104.17      40.66   2.562  0.01683 *
clippingr10   145.50      40.66   3.578  0.00145 **
clippingr5    145.33      40.66   3.574  0.00147 **
[...]
```

```
> model<-aov(biomass~clipping)
> summary.lm(model)
```

$Y_{ij} \sim N(\mu_i, \sigma^2)$, $\{Y_{ij}\}$ independants

en notant $\mu_i = \mu + \alpha_i$

Les paramètres μ_i ont des estimateurs évidents (qui sont aussi ceux des moindres carrés)

L'estimation des paramètres μ et α_i est plus problèmatique car ce modèle n'est pas identifiable
Il faut appliquer une contrainte pour pouvoir les estimer
Sous R par defaut la contrainte retenue est $\alpha_1 = 0$.
Elle aboutit aux estimateurs

$Y_{ij} \sim N(\mu_i, \sigma^2)$, $\{Y_{ij}\}$ independants
en notant $\mu_i = \mu + \alpha_i$

Les estimation de α_i s'interpretent comme des écarts à un groupe de référence qui est choisi arbitrairement comme étant le premier (par ordre alphabetique).

Estimation des paramètres et test sur les paramètres

```
> model<-lm(biomass~clipping)
> summary(model)
Call:
lm(formula = biomass ~ clipping)

Residuals:
    Min      1Q  Median      3Q     Max
-103.333  -49.667    3.417   43.375  177.667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   465.17      28.75  16.177  9.4e-15 ***
clippingn25    88.17      40.66   2.168  0.03987 *
clippingn50   104.17      40.66   2.562  0.01683 *
clippingr10   145.50      40.66   3.578  0.00145 **
clippingr5    145.33      40.66   3.574  0.00147 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70.43 on 25 degrees of freedom
Multiple R-Squared: 0.4077,     Adjusted R-squared: 0.3129
F-statistic: 4.302 on 4 and 25 DF,  p-value: 0.008752
```

Intercept est la moyenne des biomasse du traitement control (le premier dans l’ordre alphabétique).

$$\alpha_i = \bar{Y}_i - \bar{Y}_1$$

On ne peut pas directement déduire de cette table la comparaison entre r10 et r5 par exemple. Pour le faire relevel(clipping,ref="r10")

Pour se convaincre

```
> tapply(biomass,clipping,mean)
 control      n25      n50      r10       r5
465.1667 553.3333 569.3333 610.6667 610.5000
> tapply(biomass,clipping,mean)-mean(biomass[clipping=="control"])
  control       n25       n50       r10       r5
  0.00000  88.16667 104.16667 145.50000 145.33333
```

contrastes

par defaut R choisi « treatment contrast » mais on peux spécifier nous même les contrastes.
Par exemple, ici il peut être judicieux de comparer :

- le contrôle aux autres traitements
- les traitements shoot et root entre eux
- n25 vs n50
- r10 vs r5.

(rappel il existe p-1 contrastes orthogonaux)

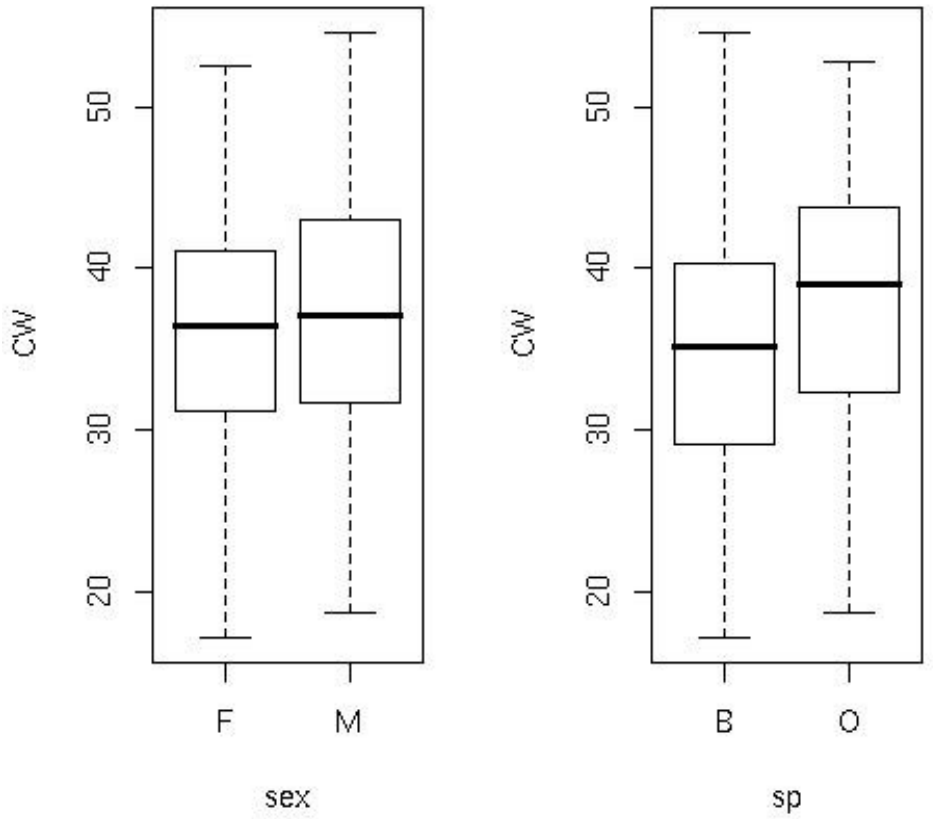
```
> contrasts(clipping)<-cbind(c(4,-1,-1,-1,-1),c(0,1,1,-1,-1),c(0,0,0,1,-1),c(0,1,-1,0,0))
> contrasts(clipping)
        [,1] [,2] [,3] [,4]
control    4    0    0    0
n25       -1    1    0    1
n50       -1    1    0   -1
r10       -1   -1    1    0
r5        -1   -1   -1    0

> model2<-lm(biomass~clipping)
> summary(model2)

[...]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 561.80000   12.85926  43.688  < 2e-16 ***
clipping1   -24.15833    6.42963  -3.757 0.000921 ***
clipping2   -24.62500   14.37708  -1.713 0.099128 .
clipping3     0.08333   20.33227   0.004 0.996762
clipping4    -8.00000   20.33227  -0.393 0.697313
[...]
```

4) prise en compte des interactions exemple de l'ANOVA à 2 facteurs

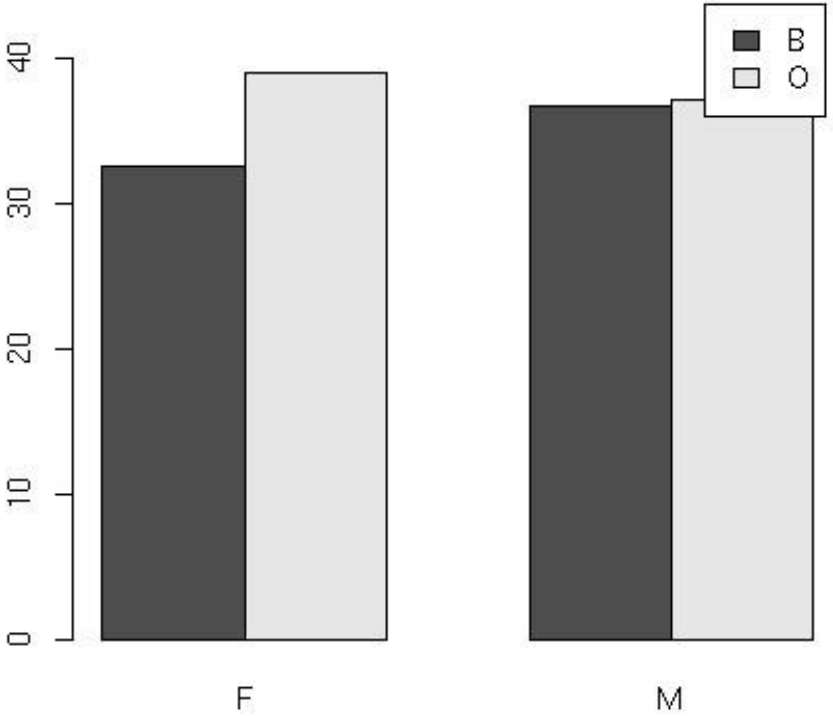
présentation

Est ce que la couleur et le sexe ont un effet sur la taille de la carapace des crabes de l'espèce Leptograpsus variegatus collecté à Fremantle, Australie ?

```
> layout(matrix(1:2,1,2))
> plot(CW~sex)
> plot(CW~sp)
```




```
>barplot(tapply(CW,list(sp,sex),mean)
+ ,ylim=c(0,45),beside=T,legend.text=
T)
```

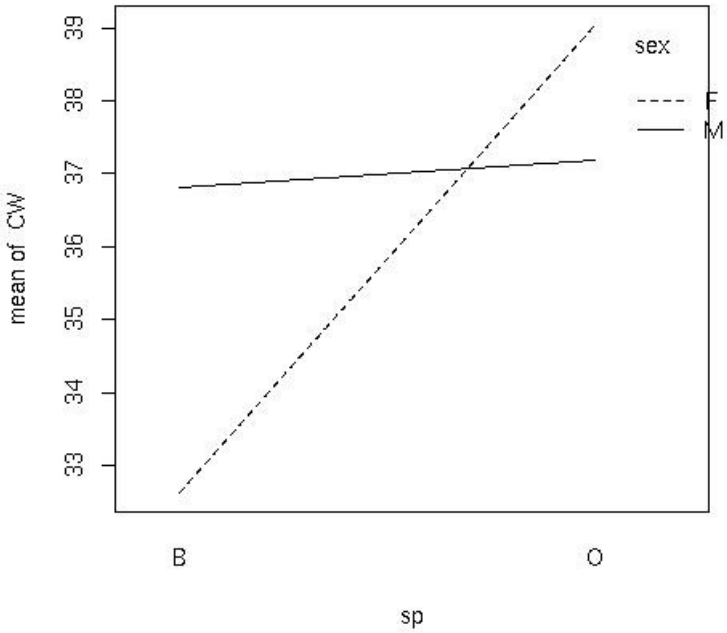



```
> tapply(CW,list(sp,sex),mean)
      F      M
B 32.624 36.810
O 39.036 37.188
```

écriture du modèle

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

> interaction.plot(sp,sex,CW)



- Variable Y_{ijk} taille de la carapace
- Le paramètre μ est un terme constant
- Le parametre α_i représente l'effet principal du sexe
- Le parametre β_j représente l'effet principal de la couleur
- Le parametre γ_{ij} est le terme d'interaction
- Variable E_{ijk} terme résiduel aléatoire

$\forall\sigma^2$ variance résiduelle

Ecriture en terme de loi des Y_{ijk}

$Y_{ijk} \sim N(\mu_{ij},\sigma^2), \{Y_{ijk}\}$ indépendants

en notant $\mu_{ij}=\mu+\alpha_i+\beta_j+\gamma_{ij}$

Sous R : lm(variable à expliquer ~ variable(s) explicative(s), ...)

```
>library(MASS)
>data(crabs)
>attach(crabs)
> model <- lm(CW~sex*sp)   ⟺   > model <- lm(CW~sex + sp + sex:sp)
```

sex:sp interaction sex/espece

Table d'analyse de la variance et test de modèles emboîtés

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

SSM=SSA+SSB+SSI

```
> model<-lm(CW~sex*sp)
> summary.aov(model)
```

somme des carrés sequentielle
(test de type I en SAS)

```
Response: CW
            Df  Sum Sq Mean Sq F value   Pr(>F)
sex          1    68.3    68.3  1.1924 0.276196
sp           1   576.3   576.3 10.0567 0.001762 **
sex:sp       1   455.1   455.1  7.9419 0.005325 **
Residuals  196 11231.8    57.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R(α/μ)

R(β/μ,α)

R(γ/μ,α,β)

Les tests des effets principaux n'ont de sens qu'en l'absence d'interaction

Estimation des paramètres et test sur les paramètres

Rappel, par defaut pour estimer les paramètres R utilise une contrainte.

$\alpha_1 = \beta_1 = \gamma_{11} = 0$

Cela revient ici à prendre la combinaison sexe feminin * couleur bleue comme reference.

```
> model<-lm(CW~sex*sp)
> summary(model)

Call:
lm(formula = CW ~ sex * sp)

Residuals:
   Min     1Q  Median     3Q    Max
-18.588  -5.294   0.151   5.335  17.790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.624      1.071  30.474  < 2e-16 ***
sexM           4.186      1.514   2.765  0.00624 **
spO            6.412      1.514   4.235 3.51e-05 ***
sexM:spO      -6.034      2.141  -2.818  0.00533 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.57 on 196 degrees of freedom
Multiple R-Squared: 0.08918,    Adjusted R-squared: 0.07524
F-statistic: 6.397 on 3 and 196 DF,  p-value: 0.0003716
```

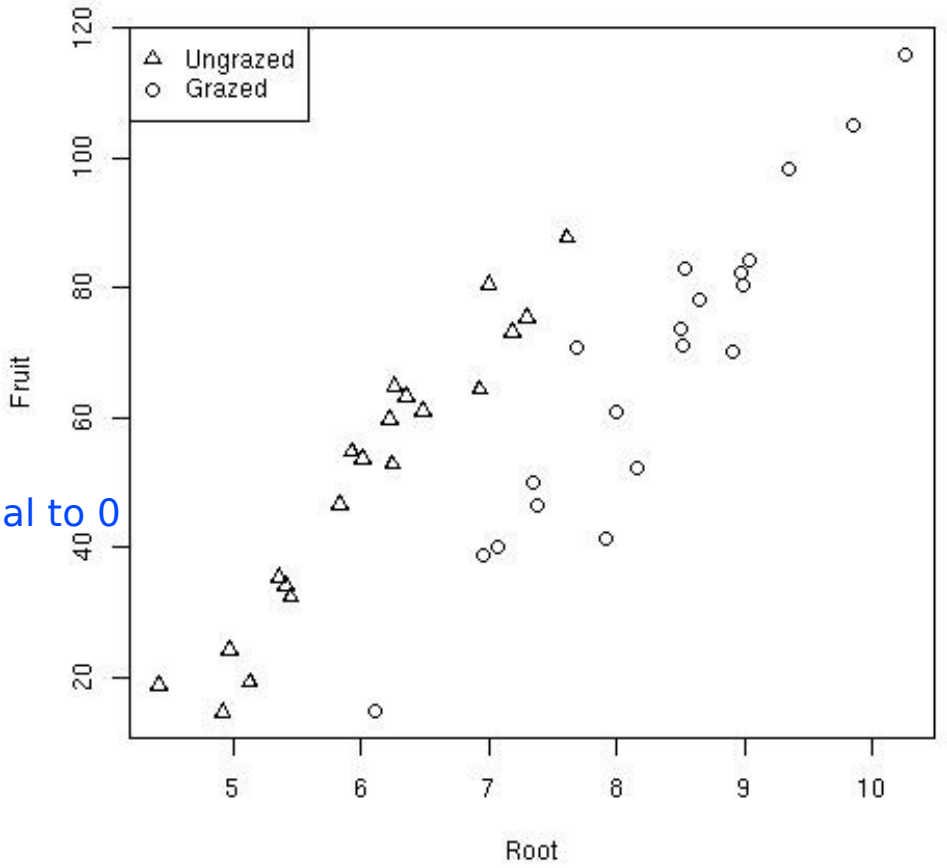
5) ANCOVA : Analyse de la covariance

présentation

Impact de l'herbivorie sur la production de graines d'une plante biannuelle dont la taille initiale varie

```
> tapply(Fruit,Grazing,mean)
  Grazed Ungrazed
 67.9405  50.8805
```

```
> t.test(Fruit~Grazing)

        Welch Two Sample t-test

data:  Fruit by Grazing
t = 2.304, df = 37.306, p-value = 0.02689
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.061464 32.058536
sample estimates:
  mean in group Grazed mean in group Ungrazed
               67.9405                 50.8805
```




```
> plot(Root,Fruit,pch=as.numeric(Grazing))
> legend("topleft",c("Ungrazed","Grazed"),pch=c(2,1))
```

Si on fait un test de comparaison de moyenne sur ces données, l'herbivorie semble **augmenter** la production de fruit ce qui est contre-intuitif ?

écriture du modèle

modèle général

$$Y_{ij} = a_i + b_i x_{ij} + E_{ij}$$

décomposition des effets

$$a_i = \mu + \alpha_i$$

$$b_i = \beta + \gamma_i$$

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \gamma_i x_{ij} + E_{ij} \qquad \{E_{ij}\} \; i.i.d. \sim N(0, \sigma^2)$$

Ecriture en terme de loi des Y_{ij}

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2), \{Y_{ij}\} \text{ independants}$$

en notant $\mu_{ij} = \mu + \alpha_i + \beta x_{ij} + \gamma_i x_{ij} + E_{ij}$

Sous R : lm(variable à expliquer ~ variable(s) explicative(s), ...)

```
>model<-lm(Fruit~Root*Grazing)
```

⟷

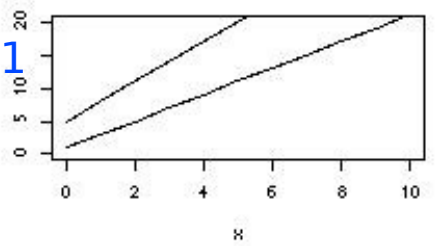
```
>model<-lm(Fruit~Root + Grazing + Root:Grazing)
```

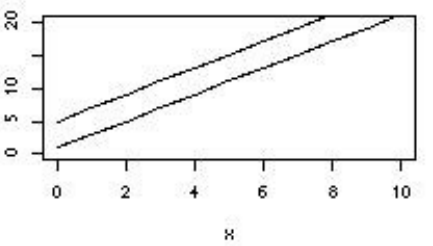
Table d'analyse de la variance et test de modèles emboîtés

Tests de modèles emboités. R(*/...)

```
> model<-lm(Fruit~Root*Grazing)
> anova(model)
Analysis of Variance Table

Response: Fruit
             Df  Sum Sq Mean Sq  F value    Pr(>F)
Root          1 16795.0 16795.0 359.9681 < 2.2e-16 ***
Grazing       1  5264.4  5264.4 112.8316 1.209e-12 ***
Root:Grazing  1     4.8     4.8   0.1031      0.75
Residuals    36  1679.6    46.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

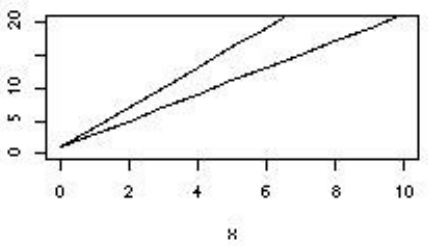
$R(\beta/\mu)$
$R(\alpha/\mu,\beta)$
$R(\gamma/\mu,\beta,\alpha)$

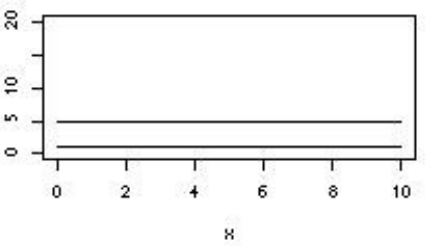


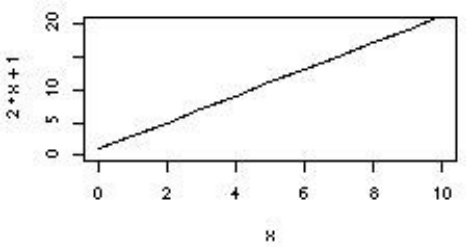


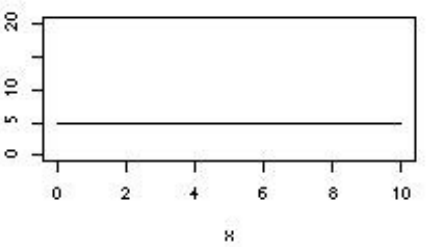
interpretation...

$$Y_{ij}=\mu+Eij$$
$$Y_{ij}=\mu+\alpha_i+Eij$$
$$Y_{ij}=\mu+\beta\,x_{ij}+Eij$$
$$Y_{ij}=\mu+\beta\,x_{ij}+\gamma_i\,x_{ij}+Eij$$
$$Y_{ij}=\mu+\alpha_i+\beta\,x_{ij}+Eij$$
$$Y_{ij}=\mu+\alpha_i+\beta\,x_{ij}+\gamma_i\,x_{ij}+Eij$$









Test des différents effets, modèle sans interactions

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + E_{ij} \qquad \{E_{ij}\} \, i.i.d. \sim N(0, \sigma^2)$$

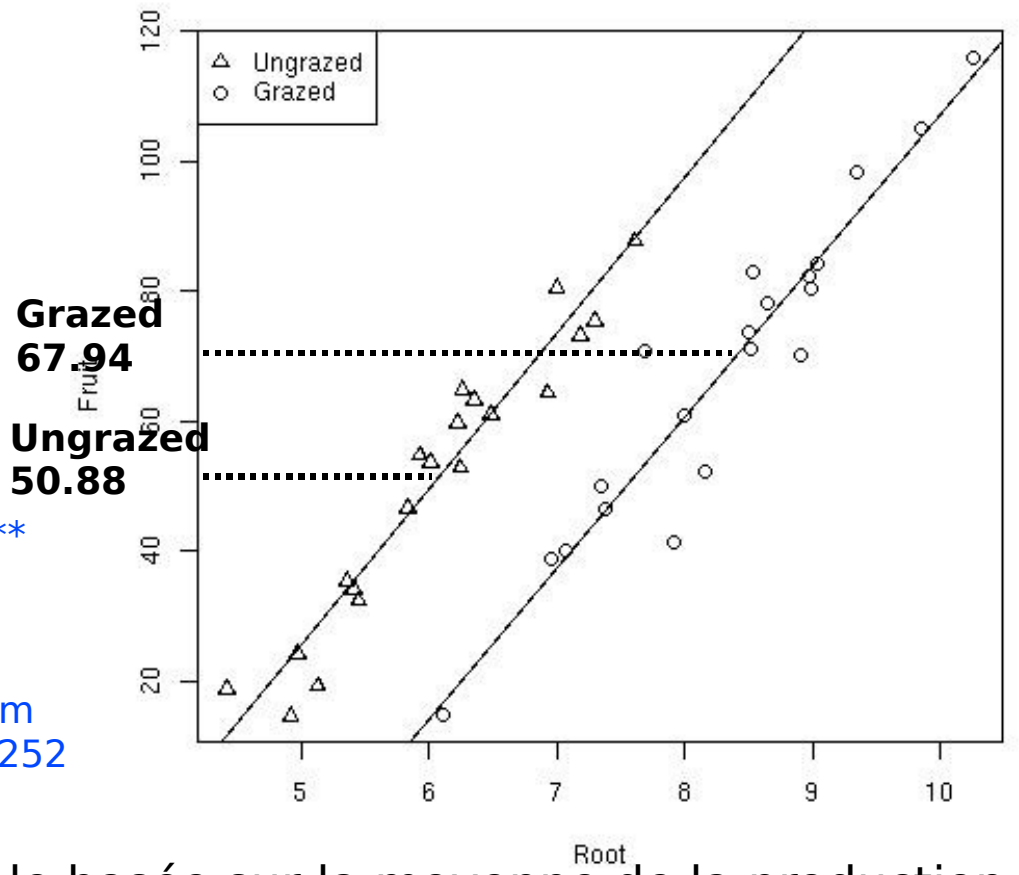
```
> model2<-update(model, ~. - Root:Grazing)
> summary.aov(model2)

            Df  Sum Sq Mean Sq F value    Pr(>F)
Root         1 16795.0 16795.0  368.91 < 2.2e-16 ***
Grazing      1  5264.4  5264.4  115.63 6.107e-13 ***
Residuals 37  1684.5    45.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$R(\beta/\mu)$

$R(\alpha/\mu,\beta)$

Estimation des paramètres et comparaisons des traitements

```
> summary(model2)
Call:
lm(formula = Fruit ~ Root + Grazing)

Residuals:
    Min      1Q  Median      3Q     Max
-17.1920  -2.8224   0.3223   3.9144  17.3290

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -127.829      9.664  -13.23 1.35e-15 ***
Root              23.560      1.149   20.51  < 2e-16 ***
GrazingUngrazed   36.103      3.357   10.75 6.11e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.747 on 37 degrees of freedom
Multiple R-Squared: 0.9291,     Adjusted R-squared: 0.9252
F-statistic: 242.3 on 2 and 37 DF,  p-value: < 2.2e-16
```



A la difference de notre interprétation initiale basée sur la moyenne de la production de graine, l'herbivorie est associée à une **réduction** de 36.103 mg de la biomasse de graines.

Analyse sur une sous-partie des données :

```
> plot(Root,Fruit,pch=as.numeric(Grazing))
> abline(lm(Fruit[Grazing=="Grazed"]~Root[Grazing=="Grazed"]))
> abline(lm(Fruit[Grazing=="Ungrazed"]~Root[Grazing=="Ungrazed"]))
> legend("topleft",c("Ungrazed","Grazed"),pch=c(2,1))
```

```
lm(Fruit[Grazing=="Grazed"]~Root[Grazing=="Grazed"])
lm(Fruit~Root, subset=(Grazing=="Grazed"))
```

6)Conclusion et perspectives

Le modèle linéaire avec R

$$y_{tj} \sim N(m_t, \sigma^2),\ \{y_{tj}\} \text{ indépendants}$$

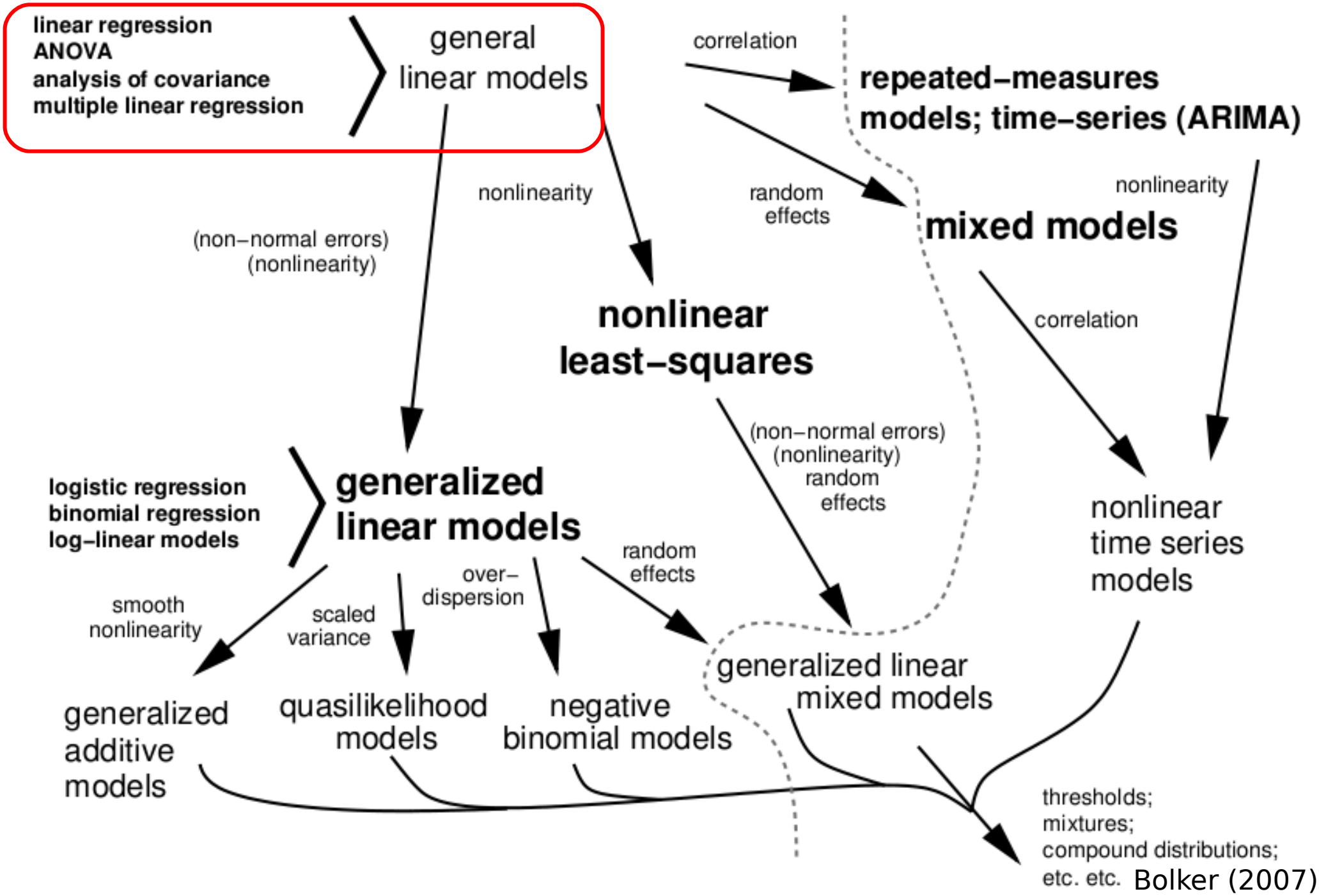
definition du modèle
> model<-lm(variable à expliquer ~ variables explicative(s), ...)

table d'analyse de la variance, test de modèles emboités
> summary.aov(model)
ou >anova(model)

Estimation des parametres et tests sur les parametres / R^2 et S^2
> summary.lm(model)
ou >summary(model)

diagnostique (validation des hypothèses du modèle)
>plot(model)

Perspectives



Bolker (2007)

References :

- The R book ; Michael J. Crawley
- Introductory Statistics With R ; Peter Dalgaard
- Le modèle linéaire ; Camille Duby
- Statistique inférentielle ; JJ. Daudin, S.Robin
- http://www.bio.ic.ac.uk/research/mjcraw/therbook/index.htm